Comparative evaluation of outcome of knee replacement operations using alternative knee prostheses.

Morris, Richard William

COMPARATIVE EVALUATION OF OUTCOME OF KNEE REPLACEMENT OPERATIONS USING ALTERNATIVE KNEE PROSTHESES

THESIS

presented for the

DEGREE

of

DOCTOR OF PHILOSOPHY

in the Faculty of Medicine

(Field of Study - Medical Statistics)

by

Richard William MORRIS

From:

Department of Public Health Medicine,

United Medical and Dental Schools of Guy's and St. Thomas' Hospitals,

University of London.

July 1993

DECLARATION OF AUTHORSHIP

This thesis has been my own work. Ms Hita Vora was employed as research assistant to the study, and was employed half time from March 1990 to August 1991. She helped to administer the questionnaire survey described in Chapters 7 to 10, and created and maintained the data base of patients included in the survey. She also entered the data on to computer from the questionnaire survey. However I carried out almost all the linking of various files and the statistical analysis described. For the early stages of this I had some assistance from Ms Katrina Man, working with me on an industrial placement from September 1990 to August 1991. Any contributions made by Ms Vora and Ms Man were under my direction.

Richard W. Morris

ACKNOWLEDGEMENTS

I should especially like to thank my supervisor, Professor W.W. Holland, for his continual support and encouragement. I am also grateful to Miss S. Chinn for encouragement and support, and for providing advice on the statistical aspects of this thesis. Ms. H. Vora carried out much administrative work for the computer survey, entered data, and maintained computer files. Ms K. Man carried out some linking of computer files and analysis under my supervision.

Ms Vora was supported as a half time research assistant for eighteen months by the Special Trustees of Guy's Hospital. I was employed in a School funded Lecturer post at UMDS whilst most of the work for this thesis was done, except that since October 1992 I have been Senior Lecturer in the Department of Public Health & Primary Care at the Royal Free Hospital School of Medicine, and I wish to thank colleagues for support during this last stage.

I am grateful to the nine orthopaedic surgeons who allowed me to approach their patients, and particularly to surgeon S7 who provided a copy of his data for the analysis in Chapter 5. I thank hospital administrative staff who helped me with finding patient records, and I thank all the patients who replied to the questionnaire survey.

The data shown in Appendix 2 from Scuderi et al (1989) were reproduced with the permission of the Journal of Bone & Joint Surgery and the authors.

Finally I thank my wife, my children, my parents and the Lord Jesus Christ for all their love and support.

CONTENTS

TABLES

FIGURES

APPENDICES

ABSTRACT

Knee replacements are increasingly carried out in developed countries and over 10,000 operations occur annually in England and Wales. Whilst the technology evolved over the last two decades many alternative prostheses were designed and orthopaedic surgeons debated over them vigorously. A formal critique of 54 papers evaluating knee prostheses published in 1987-9 revealed that most research consists of individual surgeons' case series, where comparison is problematic because of varying criteria for selection of patients and assessment of outcome.

Conventional clinical measures of outcome were rigorously investigated in a data set on 358 knee replacement patients under one surgeon. Success rates varied widely depending on alternative plausible definitions of success.

A questionnaire survey was carried out, targeted at all patients in ten series of knee replacement patients under the care of nine surgeons, involving six different prostheses. The questionnaire comprised items related to function, pain, use of aids, state of other joints and social variables, as well as seven visual analogue scales measuring patients' satisfaction. This resulted in a "knee score" which formed the chief outcome measure.

The questionnaire was subjected to tests of repeatability over one month and one year. Good agreement was found for most items, except for judging joint pain, and two functional items. The knee score demonstrated a reasonable intra class correlation, being 0.84 for two questionnaires administered at monthly intervals.

Many of the 1600 patients had died (n=296) or could not be reached at the specified address (n=138). Of the remaining 1166, 960 replied (82%). Usable data was present for 907 patients. Response was greater in patients operated upon recently and younger patients. Response rate ranged from 70% to 95% between the ten series.

The knee score varied significantly between the ten series. A prosthesis effect and surgeon effects were demonstrated. High satisfaction was associated with recently

performed operations, non use of aids, and home ownership. These did not explain the variation between the series.

Questionnaire items were compared with surgeons' recent assessments and patients were found to be markedly less optimistic, even among series with high knee scores. Clinical measures can no longer provide the only outcome in the modern era when only modest gains in success are to be expected. Patient questionnaires provide a more sensitive outcome measure. The possibilities of using these in future multi centre trials are discussed.

CHAPTER 1. OBJECTIVES AND OUTLINE

Objectives

The general objective of this thesis is to explore and apply ways of measuring outcome among groups of patients who have undergone knee replacement operations. The particular interest is in comparing outcome between groups of patients who have undergone knee replacement with different prostheses.

Background

Orthopaedic research as published in the British volume of the Journal of Bone and Joint Surgery was found to be woefully deficient in its statistical aspects (Morris 1988). Much of the research was found to be based on the reporting of uncontrolled case series, which typically led to dubious comparisons being made with other case series previously published. One of the chief difficulties appeared to be the unclear definitions of outcome variables used.

At this time there is a huge upsurge of interest in the measurement of outcomes in many branches of the health services. There is also the need for clear measurements of outcome for research purposes. Knee replacement is a medical technology which is still evolving. The numbers of knee replacement operations carried out each year has been climbing throughout the 1980s (Williams et al 1992) and is expected to continue increasing to double the present figure. Surgeons have continued to debate over the best type of prosthesis (Walker 1989) and the continued publication of case series for particular prostheses bears witness to the existence of some unresolved questions.

Methods of measuring outcome have been debated. Oglesby and Wilson (1984) pleaded that a uniform system of evaluation was more urgent than the invention of new prostheses, and Insall et al (1989) have introduced important modifications to their original system for scoring outcome (Insall et al 1976). Bellamy and Campbell (1989) have found deficiencies in many of the methods of rating outcome for knee replacement.

However most measures of outcome have been made by surgeons in outpatient clinics, and some have considered whether postal questionnaires might be a cheaper and more convenient method. Brewster and Newman (1991) were pessimistic about the value of this. However very little serious attention has been given to the idea of patient judged outcome as opposed to surgeon judged. A major part of this thesis is concerned with the development and testing of a patient questionnaire.

Outline

Chapter 2 describes the history of knee replacement and describes some of the controversies still current over choice of prosthesis. Chapter 3 presents a formal critique of 54 papers on the subject of evaluating prostheses published in 1987-9. The inadequacies of traditional methods of evaluation is demonstrated. Thus Chapter 4 discusses alternative methods of measuring outcome, including use of a postal questionnaire.

Chapter 5 presents an analysis of a data set including traditional outcome measures. The data concern some 350 patients operated upon at the Nuffield Orthopaedic Centre by a single surgeon, using a particular prosthesis. Data on later need for reoperation were available, along with clinical measures made on patients at regular follow up visits. It was thus possible to explore (1) the implications of alternative definitions of endpoint for survival analysis, (2) the changes in clinical measures over a long term follow up and whether these measures carry any prognostic value. Lastly the question of how to deal in analysis with observations made on patients who have had both knees replaced, was investigated. When these outcomes are used, long term follow up of a large number of patients is inevitable.

The postal questionnaire may be a more sensitive outcome requiring shorter follow up. Chapters 6 to 11 describe stages involved in design, conduct and analysis of a survey of patients who had undergone knee replacement. Chapter 6 describes the selection of the ten series of patients for the survey, and profiles the surgeons involved. Chapter 7 describes the development of the questionnaire and details the scheme used in its

administration. Chapter 8 displays the response rates in some detail and considers reasons for non response. Chapter 9 involves a study of the repeatability of the questionnaire when administered (i) on two occasions a month apart and (ii) on two occasions a year apart.

Chapter 10 describes the main analysis of comparing the ten series of patients concerning their reply to the questionnaire. A "knee score" is compiled which represents the chief outcome variable. This knee score is dichotomised so that patients are classed as good if their score is greater the overall median, or classed as poor otherwise. The probability of a poor result is analysed with logistic regression which compares the ten series after allowing for the effects of potentially confounding variables. Secondary comparisons were made between the four series involving the pioneer surgeons (each of whom used a different prosthesis), and between series in which different surgeons used the same prosthesis. Secondary outcome measures analysed were function score, and presence of positive and negative comments made at the end of the questionnaire.

Chapter 11 uses clinical data made available by some of the surgeons to investigate differences between surgeons' assessments of patients in outpatient clinics with the patients' responses to the questionnaire.

Chapter 12 discusses the use of a patient questionnaire in future evaluative work on the outcome of knee replacement. This chapter also considers the need and feasibility of randomised trials for further comparative research on the use of alternative prostheses.

CHAPTER 2 - THE EVOLUTION OF KNEE REPLACEMENT

Epidemiology

Arthritis is a chronic disabling condition which chiefly afflicts the elderly in its attack on the joints. In modern times it has been possible to distinguish at least two types of pathology: the most common are Osteoarthritis (OA) and Rheumatoid Arthritis (RA).

Osteoarthritis

Osteoarthritis is frequently seen as having occurred more frequently in the post-industrial era and is usually attributed to simple wear and tear of the joints. This hypothesis is currently coming under scrutiny (Dieppe 1990) but it seems fairly clear that both the prevalence and incidence of OA is greater in the older age groups. It is rarely found in persons under the age of 40 and is very common over the age of 60, though its severity varies considerably. Historically it has been defined in terms of evidence of joint space narrowing as evidenced on X rays and grading systems have been devised to show its severity (Kellgren and Lawrence 1963). The disease is more prevalent in women and this difference increases with age. Many joints are susceptible but weight bearing joints of the lower limb (in particular knees and hips), and also various hand joints, are particularly susceptible. Unfortunately OA, like RA (see below), is likely to strike at more than one site in an individual. There is a recognised syndrome of generalised OA which may partly be genetically determined, and partly by a metabolic mechanism. Site specific OA has on the other hand been associated with obesity, acute injury to the joint, or repeated impact upon the joint (seen in certain occupational groups and also joggers). Knee osteoarthritis (grade 2 or more on the Kellgren and Lawrence (1963) definition) increases from negligible before age 35 to around 30% in the over 75 age group. This increase is more dramatic in women.

Symptoms of OA including complaints of pain are related to radiologically established disease severity. Independently of this however, obese subjects and women have been reported as suffering more. In fact the relationship between signs and symptoms is disappointingly poor (Williams et al 1992) so that need for treatment is not easily

assessed.

OA is known to cause a good deal of disability, both in terms of work and daily functional activities. When striking at hips or knees, OA causes problems in getting up from a chair and climbing stairs. It has been shown furthermore that patients with OA estimated a mean of 6.7 days per month of restricted activity. Even if pain is not experienced by patients this may be due to the conscious or subconscious restriction of activity undertaken by these patients.

Published data on the natural history of OA do not give much cause for hope for newly established sufferers. Two longitudinal studies of OA of the knee showed that no patients improved and between one-third and one-half of patients deteriorated during the course of a 15 year follow up (Felson 1990). This points to the need for effective treatment programmes.

Rheumatoid arthritis

RA is not a disease which has a specific diagnostic test and thus its definition can be somewhat elusive. The American Rheumatism Association have suggested a definition for epidemiological purposes, in which at least four of seven specified criteria should apply. Other definitions have been proposed but among 25 studies of its prevalence, cited by Spector (1990), its prevalence varied surprisingly little from 1% among a wide range of global regions and ethnic groups. Its prevalence appears to increase to the age of 60, and to attain a plateau thereafter. It has been estimated that three times as many women suffer this disease as men (Gartland 1987). There has been some suggestion that it has a viral origin and that its prevalence will decline in a manner typical of a very slow epidemic (Silman 1991). It has been estimated that the life expectancy of RA patients is reduced to the tune of 3-7 years (Vandenbroucke et al 1984). However it has been suggested on the basis of available data that the incidence and severity of RA have declined over recent decades (Silman 1989).

Concerning morbidity, studies have generally shown that disease severity is progressive

and 20% will be severely incapacitated after 20 years. Unfortunately its course is variable and this causes considerable difficulties for the rheumatologist planning treatment. Some patients experience a brief illness of short duration followed by seeming recovery, some have a progressive disease punctuated by remission, and others (mainly elderly) have an aggressive form which quickly incapacitates the patient.

Compared with OA, RA is liable to strike at younger people as well as older. RA is also a disease which is more inclined to strike at peripheral joints and is less common (but by no means unknown) in the hip or shoulder. Because it is an auto-immune disease, it is not usually limited to a single joint and patients may have problems with their fingers, wrists, elbows, neck, ankles, toes and knees.

Treatment of arthritis

Conservative treatment of OA in the early stages of the disease includes simple restriction of activity, use of a walking stick and weight reduction (if appropriate). Drug treatment often consists of NSAIDs. According to Dieppe (1990) NSAID treatment is not proven to be superior to paracetamol or even placebo, and is liable to lead to side effects on stomach and kidneys.

As far as RA is concerned, progress in its treatment has been slow (Sturrock 1991) and rheumatologists have seemed to feel that all they were currently achieving was a short term suppression of the inflammation (Goddard and Butler 1984). Since RA is a disease typically of 25 years duration it would appear that no permanent solution is yet available within the rheumatology specialty. The so called second line drugs such as gold salts and penicillamine are certainly effective in the short term, but they are toxic and carry side effects such as nausea and vomiting. Newer therapies are being evaluated to see if complete as opposed to temporary remission can be achieved (Weinblatt and Maier 1991).

Surgical treatment of the knee joint.

Surgical treatment would be considered to alleviate the pain or distress not controlled by

less invasive means. Prior to 1970 two main options were available. The first was arthrodesis whereby the tibio-femoral joint was permanently fused so that no flexion of the knee could any longer take place. This had a severely limiting impact on ordinary daily activities, although it is sometimes used still if a knee replacement operation has gone badly wrong (Lettin et al 1990). A second option was osteotomy involving the cutting of parts of the bone at the joint surface. This created a secondary deformity to eliminate that caused by the arthritis, in order to redistribute load. This second option applies best when arthritis has struck just one "compartment" (side) of the knee, usually in younger patients. Proponents such as Maquet (1985) argue that for patients with osteoarthritis in one compartment only, tibial osteotomy is the treatment of choice where the medial compartment (the inside of the knee) is affected and femoral osteotomy where the lateral compartment is affected (outside of the knee). However it is felt by many that it is difficult to construct a symmetrical knee with forces evenly balanced on either side of the joint, and the procedure carries risks of complications in elderly patients (Marmor 1988). It should be pointed out that RA of the knee tends to strike the joint symmetrically and thus osteotomy is not appropriate anyway.

It should be mentioned that a more recent treatment alternative is arthroscopic debridement. This may be useful for repairing a meniscal tear in the young patient who may otherwise develop degenerative OA. It is not really seen as a serious option for more advanced disease, but it is a technology which is still evolving although its potential is for diagnostic rather than therapeutic purposes.

For RA patients, since the synovial tissue of the knee joint is destroyed, the first surgical intervention may be a synovectomy. However once joint destruction is underway, painfree motion can only be restored by something more radical.

Arthroplasty of the knee

Operations to replace artificially the knee joint have now been in use for around 40 years. Following the pioneering work of Charnley in carrying out replacement of the hip joint, a number of orthopaedic surgeons (in particular Gunston 1971) have applied

the same idea of combining metal and polyethylene materials for the development of knee joint prostheses, to be implanted in the patient with or without cement. Until the early 1970s, arthroplasty of the knee was generally regarded as still being in the experimental stage, lagging far behind hip replacements in terms of the numbers of patients treated. There was also a prevailing pessimism about the results that could be obtained (Waugh 1978) and this was born out by some early published series (Insall et al. 1976). The knee joint is known to be more complex than the "ball and socket" hip joint.

The following decade however saw a catch up and Williams et al (1992) reported around 10000 knee replacements carried out in NHS hospitals in England in the financial year 1989/90. This is currently only one third the number of total hip replacements but a catch up is expected in the next decade. In addition, results of case series published in the 1980s have been considerably more optimistic than earlier reports (Lancet 1991a). During the rapid evolution of this technology, many alternative designs have been formulated, used, and results published. For instance this author found that no less than 199 items were listed in Index Medicus 1989 under the heading "Knee prosthesis".

Much debate has occurred on both sides of the Atlantic concerning the advantages and disadvantages of alternative knee prostheses (Oglesby and Wilson 1984). In the course of two decades a number of enthusiastic consultant orthopaedic surgeons have designed prostheses (frequently named after them), have modified them, and have maintained polarised opinions. Goddard and Coleman (1992) found that 32 different prostheses were in current use in this country.

Waugh (1984) in a helpful review of the history of knee prostheses, identified four sub categories, namely (i) compartmental, (ii) condylar, (iii) semi-constrained and (iv) unconstrained. A different categorisation was given by Insall et al (1985) (i) Unicompartmental (ii) Total Condylar (a) Anatomic (b) Semi-Anatomic (iii) Constrained/ linked designs. By the late 1980s however, very few hinged replacements were being recommended and the majority of prostheses had minimal constraint (Harris and Sledge 1990). Compartmental prostheses still existed but were usually

unicompartmental (involving one side of the joint only).

The constrained "hinge" prostheses were among the earliest forms of knee prostheses. An example of a more modern form is the Stanmore knee, which is designed as a hinge, with massive femoral and tibial components. This prosthesis is constrained in the sense that the patient cannot rotate the knee (as would be the case for a normally functioning undiseased knee). Critics feel that the amount of bone which must be resected in order to achieve the insertion of such a prosthesis severely cuts down the options should the artificial knee joint later fail to work as intended (Lancet 1986b, Noble 1990). In addition the constrained nature of this knee prosthesis means that any uneven force due for example to the patient walking on uneven ground, will lead to large forces on the fixation between prosthesis and bone, leading to loosening or even breakage of the prosthesis (Waugh 1978). The natural knee joint is not a hinge so surgeons and bioengineers have searched for alternatives.

A less drastic option is found among semi constrained prostheses. Separate femoral and tibial components are attached to the distal femur and proximal tibia respectively. However in the late 1960s, a link was built into the design and each of the components had a long stem inserted along the mid-medullary canal of each bone. (Sheehan 1978, Attenborough 1978) and thus were still quite invasive.

The unconstrained prostheses are supposed to allow complete freedom in three planes: freedom of flexion (bending), of sideways movement (inwards or outwards), and of rotation. Perhaps the Oxford knee is one of the best examples, where a trough to simulate the rolling and gliding movement of the knee joint sits in between the small tibial and femoral components. Thus the femoral component, cylindrical in shape, is allowed to slide freely in the meniscal piece, to which it is geometrically matched. The tibial component has a flat surface and the meniscal piece has a flat bottom (Goodfellow and O'Connor 1978). The danger of unconstrained prostheses is that they may be just too free and in particular lead to dislocation or hyperextension. Furthermore the Oxford knee is now used on one side of the joint only (unicompartmental) when both cruciate ligaments are undamaged (Carr et al 1993b).

Compromises have been struck in the more modern forms of semi constrained prosthesis, developed and championed in particular by Insall in the USA and Freeman in Europe. These aim to resect relatively small amounts of bone, but the femoral components slide within grooves of the tibial component so that ability to rotate the knee is limited. The issues then concern how to achieve stability of the prosthesis yet allow a reasonable range of movement, how to achieve good fixation of the prosthesis to the bony tissue and thus to prevent loosening, and how to achieve correct alignment of the tibia with the femur. The chief debate now concerns whether or not the prosthesis should be able to substitute for the stabilising influence of the posterior cruciate ligament. Insall has been in favour of prostheses designed to substitute while Freeman has designed his prosthesis to allow for both options. Harris and Sledge (1990) inclined towards retention of the posterior cruciate ligament for knees with only moderate deformity, but substitution in severer cases.

These new prostheses may be termed semi constrained although being considerably less invasive than the previous generation of semi constrained prostheses. It was claimed by Freeman and Railton (1987) that up to 90% of prostheses currently implanted (including designs which do and do not substitute for the function of the posterior cruciate ligament) were of this semi constrained variety.

Walker (1989) attempted an informal meta-analysis of rates of revision operation required for different prostheses. Whilst acknowledging the problems of variability in data reporting, in rates of follow up, in criteria for revision surgery and in selection of patients, he suggested that condylar prostheses performed best. Hinged prostheses had high failure rates while for unicompartmental prostheses, appropriate selection of patients was critical to success. The latter suggestion was supported by Thornhill and Scott (1989) who gave a clearly specified list of selection criteria which excluded patients under 70, obese patients, those with reduced flexion or severe deformity and those with previous infection. If such criteria were not met, tibial osteotomies were to be preferred. Unicompartmental arthroplasty was seen as particularly demanding of surgical skill by these authors.

The general picture emerging is that less constrained prostheses are suitable for knee joints less damaged by arthritis at the time of surgery. Such prostheses will lead to greater functional capability for the patient. In severer cases, more constrained forms of prosthesis are preferable. Nonetheless, these conclusions have been culled from review papers. The actual evidence for efficacy of the various prostheses in clinical practice is examined in the next chapter.

CHAPTER 3. EVALUATION OF LITERATURE ON KNEE REPLACEMENT OPERATIONS

In order to understand the manner in which prostheses are selected by surgeons and how their continued use is justified, it was decided to survey the current literature on outcome of knee replacements. This was to clarify suspected deficiencies in the methods of evaluating prostheses.

I had already carried out an appraisal of the orthopaedic literature prior to starting the work in this thesis (Morris 1988). I had surveyed all published papers in the British volume of the Journal of Bone and Joint Surgery in 1984, and had discovered some major faults in a number of statistical aspects. It was decided in the context of this thesis to carry out a fuller analysis of studies of knee replacements in the recent literature, to assess whether the same deficiencies existed. If such deficiencies did indeed exist, then the need for new methods of evaluating prostheses would be needed. The development of such methods is the subject of the remainder of this thesis.

Summary of my previous work (Morris 1988)

All articles in the Journal of Bone and Joint Surgery (British volume) published in 1984 were examined. After excluding case reports and review articles, 103 papers were assessed. The sample of patients was adequately described in only 53 papers, the outcome variable was properly defined in only 68, and the data clearly presented in only 59. I subjectively considered that in only 35 of the papers were the conclusions justified by the results. No less than 73 papers were descriptions of case series without controls. Details of important differences in patient mix, method of assessment, type of care, and length of follow up were frequently missing. Thus proper comparisons between results presented in such papers were not actually warranted, although authors frequently wrote their conclusions as if they were. Green and Byar (1984) in listing eight categories of research study of new treatments, identified case series as the second weakest, stronger only than anecdotal case reports.

Seven of the papers which I had included in the survey referred to evaluation of knee replacement operations. My assessments of these seven papers will be referred to below.

Methods for new survey

Selection of papers

All items listed in Index Medicus 1987, 1988 and 1989 under the title "Knee Prosthesis" were investigated. Three years' worth of literature were surveyed, because this formed a short enough time span to give a snapshot of the status of publications in this field, yet long enough to accrue a sufficient number of papers. There were 129, 152 and 199 respectively in the three years. Articles not published in English were excluded, as were all articles whose title implied technical theory rather than actual results. Results which referred to revision operations of knee replacements rather than primary operations were also excluded, as were case reports. Items described as abstracts were also excluded since it would not have been possible to assess their scientific merit properly. Thus 26 papers from Index Medicus 1987 were selected, along with 39 from Index Medicus 1988, and 42 from Index Medicus 1989. The higher numbers drawn from the latter two years were mainly due to special issues of one or two journals publishing a cluster of papers on the theme of knee replacement.

The 107 different papers were published from a total of 22 different journals, 12 of which were specifically geared towards orthopaedics. Four of these journals (Clinical Orthopaedics, Journal of Arthroplasty, Journal of Bone & Joint Surgery - American volume, and Journal of Bone & Joint Surgery - British volume) contained 60 of the 107 papers. It was decided to restrict attention to these since the 60 papers were more geared towards the evaluation of outcome of knee prostheses than the rest. They were also assumed to carry more attention from the orthopaedic readership. Six were excluded because they were found upon closer examination to be case reports or reviews. Of the remaining 54 papers, 20 were from Clinical Orthopaedics, 16 from the Journal of Arthroplasty, and 18 from the Journal of Bone & Joint Surgery: eight from the American volume and 10 from the British volume.

This survey was carried out between April and July 1990.

Evaluation

The method of assessment developed by Morris (1988) was applied to the 54 papers considered here. The question concerning the adequacy of the description of treatment was omitted, since it was felt that only orthopaedic surgeons could be competent to judge this properly.

The proforma used is shown in Appendix 1.

I had found that good intra observer and inter observer agreement was obtained in my previous survey of the 103 Journal of Bone & Joint Surgery (British) (JBJS [Br]) articles. The only item on the questionnaire which showed poor agreement with other statistical observers concerned whether conclusions were justified by results presented. I had been much less inclined to consider conclusions justified than my two colleagues. I had felt that any conclusion taken from a single case series could never be justified because of the lack of suitable comparison. However in the present study it was decided that unless statements were made with inappropriate dogma, then conclusions need not be deemed unjustified.

Finally note was made for each paper concerning a number of items specific to knee replacement. Concerning the population, note was made of whether the diagnostic mix of the patients studied was stated. The actual outcome measure used in each paper was recorded. Also, each paper was examined to discover whether the sample size was stated in terms of the number of patients as well as the number of knees operated upon.

Analysis

The main purpose of this study was to be descriptive. However a number of secondary analyses were undertaken, including a comparison of various aspects of the papers between the four journals. Some of the more relevant items were compared between the

three years of publication, to investigate the possibility of a trend towards increasing statistical quality of published papers. Such comparisons would obviously have low statistical power unless the effect was dramatic. In addition it is likely that more than three years would be required to show a large difference, given that comparisons of changes over time cited by Altman (1991b) were made over periods of at least a decade. However the assessments which had previously been made on the JBJS [Br] 1984 papers were considered and the seven pertaining to knee replacement were examined separately.

Where formal comparisons were made, the chi squared test for contingency tables was used.

Results

The main findings are shown in Table 3-1, and this includes a comparison by journal in which the article was published. In the current study, it may be seen that the aim was clearly stated in the vast majority of papers. The population was properly defined in just over half the papers and the selection of the sample in nearly three quarters. As expected the majority of studies consisted of case series, and not a single randomised trial was evident although one fifth comprised some sort of trial albeit non-randomised. In contrast to my study of JBJS [Br] articles in 1984, the present study found at least one outcome variable to be properly defined in nearly every paper. The data were usually clearly presented, and statistical techniques appeared to be required for 42 papers. Only in 34 were such techniques used: in 26 the technique was described. Of the 34 papers where statistical techniques were used, only in 19 were these appropriate, in three inappropriate, and in 12 unclear (usually because the technique had not been described). It was never possible to tell whether the calculations were correct. The title seemed reasonable in all papers, but conclusions only seemed justified by results in 30 out of 54 (56%).

In three papers a statistician was coauthor, and a statistician's help was acknowledged in a further six. The sample size varied from a minimum of eight to a maximum of 8000. The median sample size was 69 with an inter quartile range of 39 to 101.

A typical case series

The profile of a typical paper is demonstrated by the following description. Seventy one patients were included for analysis and of these, 14 had undergone replacement of both knees. These 71 were a subset of an original group of 80 patients whose knees had been replaced over a five year period. Seven had died and two could not be reviewed because of severe rheumatoid arthritis. There were 62 women and nine men. The mean age at operation was 66.5 years, with a range of 27 to 79. Sixty one operations were carried out on osteoarthritis patients and 24 on rheumatoid arthritis patients. When patients were assessed, the mean time since operation was five years, with a range from three to eight years. The main outcome measure seemed to be the HSS assessment, though other outcomes such as patellar complications and radiograhic angular measurements were also mentioned. The authors declared that they had obtained excellent or good results in 90% of the operations, this being based on the number of patients who exceeded an agreed cutoff point for the HSS score. In discussion, the authors compared produced a table which compared their results with those reported in two other similar publications where a similar prosthesis was used.

Unjustified conclusions

In total, 27 reasons were noted by this assessor concerning the unjustified conclusions drawn for the 24 papers. In six papers a dubious comparison was made between results obtained for a case series with other results quoted in the literature, while in two papers the authors considered they had obtained good results without making any comparison at all. In six papers a formal statistical analysis should have been carried out (usually a significance test) to convince the reader that results obtained in one subgroup of patients was really different to another subgroup. In four papers important data were not properly presented. In three more papers it appeared that a statistical comparison of patient subgroups required some adjustment for important variables which confounded the comparison, while in three the results for two knee replacements in the same patient being treated as independent was considered to result in seriously misleading results. In two more papers non significant results from comparing outcome between

two subgroups was taken by authors as proving equivalence. Finally one paper was based on data derived from blood loss post joint replacement collected on barely one third of the sample.

Further data

There was considerable variation among the papers in the patient populations sampled, and there were more than 20 different outcome variables among the 54 papers assessed. The Hospital for Special Surgery score appeared in 18 of the papers, but this is a score used by North American rather than British authors. In all but eight papers the number of patients **and** the number of knees was given.

It was clear that some papers were geared towards evaluation for a special subpopulation of patients (for example juvenile rheumatoid arthritis).

Comparison by journal

A comparison concerning the above points was made with respect to the journal in which the papers were published. Very few differences were apparent, but this is largely due to the small number of papers published per journal. The Journal of Arthroplasty more frequently made unjustified conclusions than the rest, but this difference was not statistically significant.

Comparison by year of publication

Table 3-2 shows the numbers and percentages of papers judged to hold certain statistical attributes according to year of publication. Chi-squared tests were carried out on relevant 2 x 3 tables but no clear differences could be demonstrated except concerning whether conclusions were justified ($p=0.005$). This suggested that conclusions were increasingly likely to be deemed justified with increasing year of publication (25% for 1987 papers and 85% for 1989 papers). However this effect was not accompanied by any similar trends for other statistical attributes.

Comparison with 1984 JBJS series

With regard to the 54 papers appraised in this study, the statistical quality appeared to be improved when compared with the 103 papers in the British Journal of Bone and Joint Surgery in 1984 (Table 3-3). The sample selection was described properly in three quarters of the articles assessed here compared with half in the earlier study ($p=0.007$), and the outcome variable was nearly always described compared with only two-thirds of the 1984 papers ($p=0.0005$). Data were usually reasonably clearly presented here ($p=0.03$). Statistical techniques were more likely to be described here (34/42 (81%) compared with 24/66 (36%)), although not substantially more likely to be appropriate (19/34 (56%) compared with 11/24 (46%)). The conclusions seemed justified in over half these papers compared with one third of the 1984 papers ($p=0.015$). Somewhat less encouragingly, the population was specified here in barely half the papers, compared with nearly 90% in 1984 ($p=0.0001$).

The seven knee replacement papers published in 1984 were considered satisfactory in most aspects, though in only one were the conclusions deemed justified. It had been considered that all seven papers required statistical analysis, but only three had done any. The required analysis was usually the calculation of a confidence interval. Those ten knee replacement papers published in 1987-9 in JBJS [Br] seemed similar in quality to those published in 1984. In 1984 conclusions were less likely to be considered justified but the population was more likely to have been properly defined. Such differences could, of course, very easily have arisen by chance.

Discussion

This survey of papers concerned with the evaluation of knee replacement operations was restricted to 60 out of 107 papers coming from 4 out of 22 journals. Thus smaller orthopaedic journals and non orthopaedic journals were excluded from this assessment. It was considered that Clinical Orthopaedics and the American and British Journals of Bone and Joint Surgery had the widest appeal among orthopaedic surgeons, and that papers published here carried substantial authority in the

orthopaedic community. Thus diffusion of expertise concerning knee replacement surgery is likely to emanate from these sources. In addition, the Journal of Arthroplasty is probably the most specialised journal for joint replacements in the world, though it is a younger journal (started in 1986), presumably with a narrower circulation. If good or bad research is published in these four journals, this will be reflected in the appropriate or inappropriate use of certain prostheses in current and future clinical practice.

Some evidence has been found that statistical quality of these 54 knee replacement papers was higher than that for the publications in JBJS [Br] in 1984. There are several possible explanations. Firstly, there may have been a rising consciousness of the need for proper scientific evaluation which could have applied to all branches of orthopaedic research, not merely concerning knee replacement. No evidence for an improvement was found when eight papers on knee replacements published in JBJS [Br] (1984) were compared with the ten published in the same journal in 1987-9, though the numbers involved are really too small to show anything but a dramatic improvement. It did however seem that within the present series, unjustified conclusions were less likely among articles published in 1989 than in 1987 or 1988.

It is also possible that the evaluation of knee replacement operations brings a special need for a carefully thought out quantitative study. Since it is now a well established treatment (over 10000 operations carried out in NHS hospitals in England in the financial year 1989-90: Williams et al 1992), research has moved well beyond the anecdotal in this area, and the desire to establish improvements in detailed aspects of treatment must lead to proper quantitative methodology. Thus any improvement between the 103 papers in JBJS [Br] 1984 and all knee replacement papers in 1987 to 1989 may be due to the subject matter rather than a real time trend.

Finally however the possibility must be acknowledged that assessment of statistical quality of research is subjective, and that the apparent improvement may be due to this assessor relaxing his criteria.

Nonetheless the possible improvements seen among this series of knee replacement papers cannot hide the remaining problems. These were the poor definition of the population and sample selection, the poor data presentation and inappropriate use of statistics.

Criticism of the typical case series

The 71 patients reviewed were clearly asked to come for a clinical assessment all at once. Nine patients could not come because they had died or their arthritis was too severe. It is not possible to say whether the outcome for these nine would have spoiled the results presented; it is unlikely that the two with severe arthritis would have given good or excellent results. Fourteen patients underwent replacement of both knees, and in analysis the two observations for each of these patients were regarded as if they were statistically independent. The gender distribution was rightly presented according to patient, but the diagnostic distribution was incorrectly presented according to the knee. This is incorrect because a patient would not have a knee replacement for osteoarthritis in one knee and rheumatoid arthritis in the other. It is not clear how whether the age distribution was presented by patient or by knee, but in any case the range is a poor measure of the variation. It is well known that the range increases with the sample size, and is susceptible to single outlying values. The same problem occurred concerning length of time since operation.

The worrying aspect of the paper was its comparison with two other case series. The difference in length of follow up and percentage of osteoarthritis cases was mentioned, but the possible effect of this on results is not even discussed. Mention is made of the lower preoperative HSS scores in the authors' series. Lesser knee flexion was obtained in the authors' series, and this was noted but not discussed. The possibility that such a difference could have occurred by chance was not explored. The comparison was made wholly uncritically, and the authors ended their summary by declaring the prosthesis useful for patients with a severely deformed joint.

Population definition

For the results of a study of knee replacements to have any worth in clinical practice, it must be clear to whom the results can be generalised. It is likely that for many series of patients under the care of a particular consultant, the definition is unknown. Criteria for who should and who should not receive a knee replacement may well vary between surgeons, depending on the patients' age, social circumstances, fitness for general anaesthetic, appearance of their X rays and other aspects which are not easily defined. Indeed the population of patients who confront a consultant orthopaedic surgeon with the possibility of knee replacement may have already been selected by a referral pattern involving rheumatologists and general practitioners. This problem, though not necessarily the fault of orthopaedic surgeons who publish their results, may still severely hamper comparisons of case series.

Some papers restricted their attention to well defined subgroups (eg juvenile rheumatoid arthritis), and one paper was particularly clear, including patients with osteoarthritis in at least two compartments of the knee where deformity was less than 35 degrees and flexion contracture less than 30 degrees.

Sample selection

This relates to the problem of population definition; but there is in addition a responsibility for authors to say how many patients were operated upon over a particular period of time, and out of these how many have been reviewed, how many have died, how many could not be traced, how many refused to attend a clinic for assessment and so on. This is particularly relevant because those who are actually included in analysis are likely to be a biased subset of those on whom knee replacement was carried out; indeed if the outcome consists of an assessment by a surgeon in an outpatient clinic they have at least to be fit enough to make the journey to the clinic. Thus results are liable to be biased towards optimism. Authors could give some information to compare the patients included in analysis with those who had the operation but are not included. Thus statements such as "This review concerns twenty patients who underwent knee

replacement in 1983 and were assessed in 1988" are inadequate.

This point is further illustrated by a major analysis of 8000 knee replacements carried out in Sweden (Knutson et al. 1986), involving replacements in 4505 knees where osteoarthritis (OA) was the diagnosis and 3495 knees with rheumatoid arthritis (RA). During the six years of follow up deaths occurred for 498 of the former group and 653 of the latter (11% and 19% respectively, check). In addition 4.4% of OA cases were lost to follow up compared with 5.6% of RA cases. Thus patients in whom a six year review were conducted were weighted towards OA cases. Any estimate of success rates based on the six year review would be overoptimistic if the procedure was actually better for OA patients than RA patients. This shows the need for describing and comparing patients followed up with those not followed up.

Poor data presentation

Results in papers should be presented in a way that readers can easily follow conclusions drawn. While statistical analysis is naturally important, it should not be described in such a way that the reader is thwarted from inspecting raw data. The Acta Scandinavica Orthopedica (1988) declared a policy of publishing the full data set. Whilst this may not be practical for large data sets, authors should provide clear tables and figures so that analysis may be seen to support these.

The important background information was somewhat patchily presented in the 54 knee replacement papers. Not all papers stated the number of patients **and** the number of knees. Even those that did tended to present the distribution of patient characteristics such as type of arthritis according to the number of knees.

The use of the range for continuous variables such as age and length of follow up should be supplemented, if not replaced, by use of the standard deviation or interquartile range.

Inappropriate statistical analysis

A problem found in eight papers was that authors needed to carry out statistical analysis yet did not do so. On some occasions the clear presentation of raw data may allow authors to omit formal analysis since the conclusion is so obvious. More often however differences in outcome between subgroups in some postoperative measure are not so clear cut and need to be tested to check whether they are simply due to chance.

Some authors carried out analysis but did not make it clear (at least to this assessor) exactly what they had done. Altman (1991a) suggested that statistical methods should be described in such a way that if readers had access to the raw data, they could reproduce what the authors had done. Some authors showed lack of understanding of the appropriateness of paired and unpaired t tests, or even of the difference between t tests and chi-squared tests.

Two situations emerged where some degree of statistical sophistication was evidently required. Firstly authors sometimes needed a multifactorial analysis to allow for the effects of numerous factors on outcome. Outcome in knee replacement may be influenced by the patients' type of disease, age at operation, gender, preoperative status of knee and other medical problems as well as the type of treatment used. Sometimes treatment comparisons may not be made without some analysis which allows for the effects of all these other variables.

A second problem concerned the use of two knee replacements in a single patient as if they were statistically independent. All significance tests assume independence of observations. This cannot be the case in this situation for two reasons:

(i) Certain patient characteristics will lead to a degree of similarity in outcome for both knees.

(ii) The effect of success (or failure) in one replaced knee will have a beneficial (or

detrimental) effect on the other.

The appropriate handling of such data sets will be explored in more depth elsewhere in this thesis (Chapter 5). A simple approach for future authors would be to include only the first knee replacement in patients who have had both knees replaced. Whilst it may be objected that this approach is wasteful of data, regarding all operations as independent could be regarded as overuse. Morris (1993b) has explored this problem for any randomised trial in orthopaedics where individual patients may contribute to the data more than once.

Comparability of results presented

Little effort was apparent from authors of these papers at achieving uniformity in measuring outcome. One outcome which was used in 18 of the 54 papers was the HSS, which has recently been modified (Insall et al 1989). However no study of inter observer variability has been reported even for the HSS. This places considerable questions against the validity of comparing case series with each other.

Even where the British Association for Surgery of the Knee measure was used (Aichroth et al 1978) in five separate case series published together in JBJS(B) in 1978, strong objections were made when it was noted that the patient populations differed in important background factors (Tew and Waugh 1979).

Recommendations

Some of the remedies for the deficiencies outlined have been published by Carr et al (1993a). This paper was partly driven by the findings reported above. Carr et al provide a "joint replacement trial checklist". This consists of seven guidelines for planning a trial, five for performing it, six for analysing it and four for interpreting the results.

Conclusion

Despite some encouraging signs in the statistical quality of the 54 knee replacement papers reviewed, there are a number of fundamental criticisms which place in doubt some of the conclusions drawn by authors of these papers. Relevant data were sometimes omitted and sometimes presented and analysed inappropriately. Systematic formal comparisons of outcome between groups of patients were usually lacking or were not well performed. Case series have often been used to draw inferences which cannot be warranted concerning the efficacy of certain prostheses. Based on the recommendations of Carr et al, radical changes are required to the methods traditionally used for evaluating joint prostheses.

TABLE 3-1

Quality of published papers on knee replacement published 1987-9 by journal

Number of papers:

Aspects of papers	Clin Orth	J Arth	JBJS [Am]	JBJS [Br]	Total
Study aim described?					
Yes	19	13	8	10	50
No	1	3	0	0	4
Population defined?					
Yes	11	12	5	3	31
No	9	4	3	7	23
Sample described?					
Yes	16	11	6	7	40
No	4	5	2	3	14
Type of study ?					
Case series	14	8	4	8	34
Case control	0	4	1	0	5
Concurrent non RCT	3	3	3	2	11
RCT	0	0	0	0	0
Other	3	1	0	0	4

Quality of published papers on knee replacement published 1987-9 by journal

Number of papers:

Aspects of papers	Clin Orth	J Arth	JBJS [Am]	JBJS [Br]	Total
Outcome defined?					
Yes	18	15	8	9	50
No	2	1	0	1	4
Data presented?					
Yes	15	12	5	9	41
No	5	4	3	1	13
Analysis needed?					
Yes	14	14	7	7	42
No	6	2	1	3	12
Statistics used?					
Yes	11	12	5	6	34
No	3	2	2	1	8
Not applicable	6	2	1	3	12
Statistics described?					
Yes	9	8	5	4	26
No	2	4	0	2	8
Not applicable	9	4	3	4	20

Quality of published papers on knee replacement published 1987-9 by journal

Number of papers:

Aspects of papers	Clin Orth	J Arth	JBJS [Am]	JBJS [Br]	Total
Statistics appropriate?					
Yes	8	5	2	4	19
No	0	2	1	0	3
Unclear	3	5	2	2	12
Not applicable	9	4	3	4	20
Title appropriate?					
Yes	20	16	8	10	54
Conclusion appropriate?					
Yes	13	6	4	7	30
No	7	10	4	3	24

TABLE 3-2

Comparison of quality of published papers on knee replacement by year of publication

Number (%) for which answer was yes:

Aspects of papers	1987 (n=16)	1988 (n=25)	1989 (n=13)
Study aim described?	15(94)	23(92)	12(92)
Population defined?	10(63)	14(56)	7(54)
Sample described?	10(63)	20(80)	10(77)
Type of study : case series?	8(50)	19(76)	7(54)
Outcome defined?	15(94)	25(100)	10(77)
Data presented?	12(75)	20(80)	9(69)
Conclusions justified?	4(25)	15(60)	11(85)

TABLE 3-3

Comparison of quality of published papers : Knee papers published 1987-9 versus all JBJS [Br] 1984 papers.

Number (%) for which answer was yes:

		JBJS [Br] 1984	
Aspects of papers	Knee papers (n=54)	All papers (n=103)	Knee papers (n=7)
Study aim described?	50(93)	99(96)	7(100)
Population defined?	31(57)	89(86)	6(86)
Sample described?	40(74)	52(51)	6(86)
Type of study : case series?	34(63)	73(71)	6(86)
Outcome defined?	50(93)	68(66)	5(71)
Data presented?	41(76)	59(57)	6(86)
Conclusions justified?	30(56)	35(34)	1(14)

CHAPTER 4. ALTERNATIVE METHODS OF MEASURING OUTCOME

The previous chapter showed that a particular deficiency of research on knee replacements was the inadequate definition of the outcome variable. In addressing this problem, the options available in measuring the outcome of knee replacement need to be considered.

Long term clinical outcome

Since joint replacement involves a major piece of surgery involving a hospital stay of two or three weeks and a long period of rehabilitation, a natural question concerns how long the implanted prosthesis will function before further invasive surgery becomes necessary. To assess this, a particularly popular statistical technique has been survival analysis, where the endpoint is "revision surgery". This was seriously introduced as a means of evaluating longevity of joint replacements by Dobbs (1980) and many authors have since employed this technique. However it has a number of limitations and problems which have been described by Carr et al (1993a) and are as follows:-

1. Long term follow up

Surgeons who are committed to a regular monitoring of their patients' progress will typically schedule follow up appointments six months and a year after surgery and yearly thereafter. The logistics for doing this are heavily dependent upon staff resources in the orthopaedic department and motivation by the patient. Inevitably not all patients continue to be followed up as scheduled and such patients are deemed "lost to follow up". In a survival analysis they are assumed to have had the same risk of needing further surgery as those who continue to be followed up. This assumption has been questioned since it has been suspected that non compliance by patients with follow up appointments may imply they are seeking attention for problems with their replaced joint from other surgeons.

Dorey and Amstutz (1989) in their study of 100 hip replacements suggested these fears

may have little ground. They made vigorous attempts to follow up 45 defaulters who were lacking complete 8 year follow up data, and were successful with 35 of these. Their recompiled survival curve differed little from that compiled before the vigorous follow up of defaulters. A similar study was carried out in France by Postel et al (1987) on larger numbers of hip replacements with similar conclusions. Once again however around 10% of patients could not be traced despite vigorous efforts, and Laupacis (1989) argued that such patients may be the very ones whose survival experience would genuinely differ from the rest.

Indeed high survival rates reported after 10-15 years of patient follow up (Ranawat et al 1988, Scuderi et al 1989, Gill and Mills 1991) may be seriously spoiled if some of those patients with incomplete follow up were in fact failures. Scuderi et al (1989) included a table of failures and withdrawals of 224 knee replacements carried out between 1974 and 1978 (shown in Appendix 2). If five per cent of the withdrawals within each particular year of follow up were regarded as failures, the life table may be recalculated, and it is found that the 10 year survival rate drops from 91% to 87%. If 10 per cent of these withdrawals are regarded as failures, the 10 year survival rate drops further to 81%. In the context of recent publications, such results would be regarded as very indifferent (Noble and Hilton 1991). Figure 4-1 shows the full impact of these fairly mild assumptions on the results of Scuderi et al.

2. Death

It has been suggested that reliable survival rates may only be calculated at time points when at least 50 patients have been followed up (Lettin et al 1991). Since orthopaedic surgeons are treating an elderly population, many patients would be expected to die within a ten year follow up period. All published survival analyses of joint replacements have regarded deaths in the same way as losses to follow up, so that those who died were assumed to be at the same risk of needing further surgery for their joint replacement as patients who lived. The only evidence for this was presented by Postel et al (1987) for their hip replacement series. They found no significant difference in hip function between those who subsequently died and those who subsequently lived, at matched time points

of postoperative follow up.

3. The definition of failure

As was alluded to, this concerns the decision to carry out revision surgery. Though this seems a hard endpoint, decisions to carry out revision surgery vary between surgeons. Such decisions are also dependent on the age and medical fitness of patients. In terms of comparing prostheses, thresholds for revision surgery may vary according to type of prosthesis. Tew and Waugh (1982) introduced the concept of "pain free" survival of prostheses. Apart from the definition of whether pain is experienced or not, it is an entity which can appear and disappear. This means it does not fit easily into the framework of survival analysis where once subjects reach the endpoint, they are unable to emerge from it again.

4. Small numbers and statistical power

If the spectacular survival rates published in recent years are to be taken at face value, survival analysis may no longer be sufficiently sensitive to demonstrate differences between two groups of patients treated with different prostheses. Long term follow up would be required on thousands of patients (as seen for example in the 9200 knee replacements reported by Rand and Ilstrup 1991) to accrue sufficient numbers of patients who reached the defined endpoint.

5. Obsolescence of results

The dual requirement of large numbers of patients and long term follow up may well mean that results may not be publishable until 20 years after the introduction of a new prosthesis. By this time many new prostheses will have been introduced.

Some of the problems listed above were investigated empirically using data on the long term follow up of knee replacements carried out on over 350 patients by one surgeon. This will be reported in Chapter 5.

Short term outcome

Because of the difficulties of following an elderly population over a long period, searches have been made for a satisfactory measure of outcome which would be rapidly available (say within two or three years of operation) and make it possible to evaluate formally a new knee prosthesis soon after its introduction.

A promising outcome was found on the Denham knee series in the radiographic measurement of early postoperative alignment of the knee. This measure was found to be predictive of loosening of the joint at eight years post-operation, in a manner independent of preoperative deformity (Jeffery et al 1991). However this result arises from a prosthesis designed especially to foster good postoperative alignment. There are plenty of other prostheses on the market where alignment is not seen as a primary goal, and postoperative alignment may not relate to long term success in the same manner. Tew and Waugh (1985) whilst finding some relationship between malalignment and failure in 428 patients using six different prostheses, felt that malalignment was by no means the most important cause of failure.

Albrektsson et al (1990) have made measures using the roentgen-stereophotogrammetric (RSA) technique and have thus been able to detect minor suggestions of loosening of a prosthesis at an early stage. It was not clear however, whether such apparent signs of loosening would actually lead to clinically problematic loosening at a later stage.

A first attempt to address this problem was made by Grewal et al (1992) on the same clinical material. The mean migration in the first postoperative year for each of three subgroups of prosthesis was calculated, along with the five year survival rates of the same subgroups. The ranking of the three mean migration values was the same as for the five year failure rates. However the mean migration rates were calculated on small subsets of the patients used for survival analysis. The statistical technique is thus based on a rank correlation on a sample size of three. For each of the three observations the values of the two variables being compared is derived from differently aggregated groups of patients. Though the data are at least consistent with the idea of migration being a

predictor of later failure, they can hardly provide robust scientific evidence. Nor do the data give any idea of how great the predictive value of migration would be.

Without a long term follow up study, no-one will know whether RSA (which involves expensive technology) is of any long term predictive value. In addition this technique may only be useful or even applicable for the prosthesis reported by Grewal et al.

The problem of whether an early postoperative measure could predict long term success would also apply to any simple clinical assessment of patients made in the early postoperative period. A long term follow up is needed of the predictive value of assessments by either the surgeon or patient of pain, degree of deformity, or mobility. It could be that an artificial knee joint which fails after 10 years appears highly satisfactory for at least 9 years on a superficial clinical assessment. Since data were available to this researcher from one particular surgeon with a long and regular follow up, the search was made for indices which may be assessed soon after operation and predict later success or failure of the prosthesis. The results of this analysis are shown in Chapter 5.

Patient centred outcome

Much of the research literature on the outcome of knee replacements has at best paid lip service to patient related measures of outcome. However the Patient Outcome Research Team (Indiana, USA), concerned with assessing and improving outcome in knee replacement and other medical technologies, have recognised that surveys of patients' function and pain are important (Freund et al 1990). Evaluation of joint replacement in general suggests it ranks highly (Williams 1991), with the cost per QALY a small fraction of the cost of organ transplant procedures and bypass grafting. Similarly Liang et al (1986) found that patients with hip and knee replacements had very high "cost effectiveness" indices with regard to improvement in well being. More recently a study of hip and knee replacements at the London Hospital found that both were associated with marked improvements in pain and disability (Kirwan et al 1992). Knee replacement itself would seem to have won a vote of confidence from patients, although Drewett et al

(1992) calculated QALYs for 26 knee replacement patients and found markedly lower ratings than those reported by Williams (1991).

Black et al (1991) have written, "Patients' opinions may be more important than those of doctors when such improvements in quality of life as symptom relief are being studied". Surgeons have certainly asked their patients' opinions , but usually only as an appendage to an overall clinical assessment (Aichroth et al 1978). However researchers have acknowledged the "gratitude factor" (Thacker and Fulford 1986) which would naturally operate most powerfully if the consultant who carried out the operation also does the assessment. This difficulty is mitigated by consultants who ask their registrar or senior registrar to carry out follow up assessments. Since such personnel change rapidly however, it does not seem realistic to expect consistent criteria to apply to a large number of patients assessed over a long period. Moreover objectivity can scarcely be guaranteed when the task is delegated to a more junior member of the surgical team (Apley 1990) . Comparisons between case series where use of particular prostheses are reported, cannot be seen as reliable if patient satisfaction was assessed in an outpatient clinic.

It was thus felt that a postal questionnaire, sent directly to the patients' home address, and to be returned to a researcher with no involvement in clinical care for the patient, would be likely to elicit less biased responses concerning the true effectiveness of the prosthesis in the patient's daily life.

The use of patient questionnaires has been considered in three studies of surgical patients and these are reviewed below.

Carlson and Pelletieri (1989)

Ninety patients treated for various neurosurgical conditions in Uppsala, Sweden, were investigated. On the day of discharge, each patient and responsible physician was supplied with a questionnaire in which each party graded the result of treatment as very good, good, fair, bad or very bad. A similar exercise was carried out at follow up 8 to

24 months after discharge. At discharge 75 out of 90 (83%) patients agreed with their physician, with disagreements equally likely to show optimism or pessimism by the physician relative to the patient. At follow up only 40 out of 69 (58%) agreements were found; among the 29 disagreements, 21 consisted of relative optimism by the physician. Disagreements were more likely among patients with painful or degenerative conditions.

Black et al (1991)

Three hundred and eighty eight patients undergoing Transurethral Resection of the Prostate under the care of 25 different surgeons in two British Regional Health Authorities were studied. They completed a questionnaire before surgery and at 3, 6 and 12 months afterwards.Patients were randomly assigned to attend follow up appointments at 3, 6 or 12 months after the operation, when surgeons assessed the patients' status. Agreement was assessed in terms of the presence or absence of symptoms, comorbidities and complications. Agreement rates were particularly good (over 90%) in reports of past medical history (for instance, occurrence of stroke or myocardial infarction) and specific postoperative complications. Agreement was less impressive (between 50% and 70%) for postoperative symptoms. Where disagreements occurred, there were more situations of patients considering the situation worse than the surgeon for six out of seven symptoms. The differences are in fact statistically significant using McNemar's test for five of these. No factors could be related to likelihood of disagreement.

Brewster and Newman (1991)

Seventy three knee replacement patients (20 before operation and 53 afterwards) were given a questionnaire to complete in outpatient clinics before a physiotherapist made an assessment.The patient's form and therapist's form were designed to be as similar as possible, following the Bristol Knee Scoring form which had been designed for routine use by physiotherapists several years previously. Therapists' and patients' assessments were compared by categorising agreement as "reasonable" (scores being within 10% of each other), "patient's score higher" or "patient's score lower". For the total score, reasonable agreement was found for only 25 out of 73 (34%). In 42 cases the patient

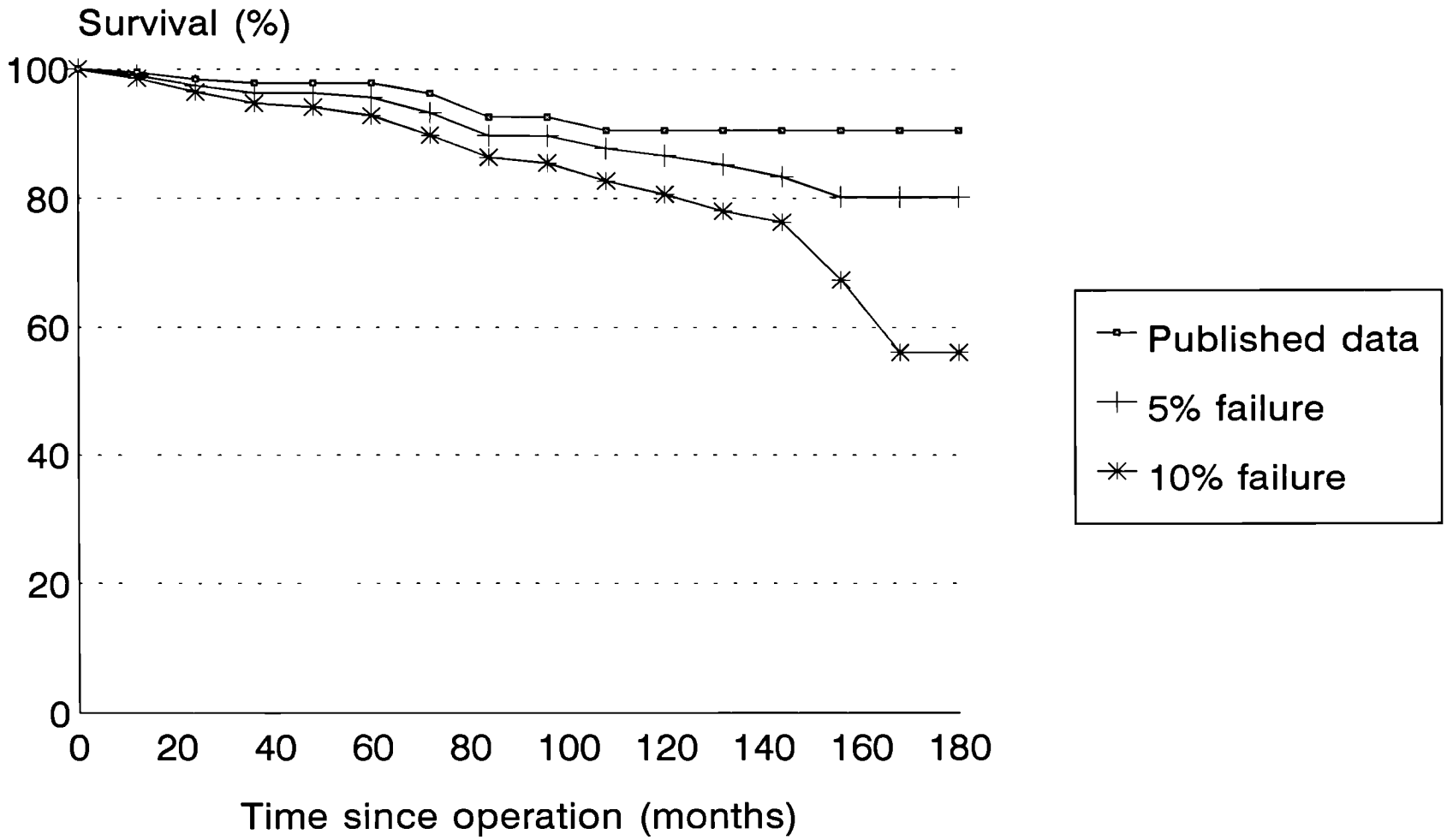
scored lower and in 6 cases the patient scored higher. This bias was particularly marked for postoperative patients. It was concluded that patients could not reliably assess outcome of the replacement of the knee joints and thus postal questionnaires should be viewed with caution.

All three studies are consistent in finding patients more pessimistic than doctors. The same trend was found by Jachuk et al (1982) who studied patients on hypotensive medication; all patients were considered better by their GP but only half the patients thought this, while most of the patients' relatives considered them worse.

Brewster and Newman whose study is most relevant to this thesis, considered the lack of agreement as serious. Their conclusion that the disagreement was the fault of the patients has been challenged by Pound et al (1992) and Streetly and Morris (1992). There seems no good reason to conclude that patients' assessments were less reliable than those of the surgeon when subjective symptoms were being considered.

Although Brewster and Newman applied their findings to the use of a postal questionnaire, no formal testing of a real patient questionnaire has yet taken place for knee replacement patients. The main purpose of this thesis was to test measures arising from a postal questionnaire in order to compare knee replacement patients treated with different prostheses, and results will be reported in Chapter 10. Comparisons between responses to the questionnaire with clinical measures of outcome made in follow up clinics are shown in Chapter 11. The reliability of this postal questionnaire was first examined and this is reported in Chapter 8.

Figure 4-1. Survival of Total Condylar Prosthesis under different assumptions



CHAPTER 5. CONVENTIONAL CLINICAL OUTCOME MEASURES - A STUDY OF THE OXFORD KNEE REPLACEMENT SERIES

Introduction

Data collected by one surgeon among the 10 series reported in this thesis were made available to RWM. This surgeon had carried out knee replacements on over 350 patients from the mid 1970s until 1990. He or his registrars had assessed the progress of each patient's replaced knee (or knees) at regular follow up clinics. Data were recorded from these assessments and this was linked to the occurrence of further operations on the replaced knee during the period of follow up. Since these data involve a longitudinal study, collected in order to form a clear idea of the long term outcome of the Oxford knee prosthesis, certain questions may be addressed which are not possible within the constraints of the cross sectional study reported in Chapter 10 of this thesis.

This form of data collection is traditional in assessing long term outcome of joint replacements. Assessments are made in hospital outpatient clinics by surgeons. The same principle of data collection is practised by a number of other orthopaedic surgeons for their particular series of knee replacement operations, but the Oxford data are hard to match for completeness. The data have been put to two main uses to date. Firstly overall "survival" of the Oxford knee joint has been assessed along with outcome in terms of pain relief, function and joint movement. This has been done for the bicompartmental form (Goodfellow and O'Connor 1986) and the unicompartmental form (Carr et al 1993b). Secondly factors related to the risk of failure of the implanted prosthesis have been identified (Goodfellow et al 1988). Further questions were explored however for the purposes of this thesis.

1. How should the endpoint be defined for survival analysis? The conventional definition relates to the point at which revision surgery is undertaken (Dobbs 1980). This has been contentious for a long period (Tew and Waugh 1982, Nellisen et al 1992). One particular issue has concerned patients who experience pain yet are deemed medically unfit for revision surgery.

2. Do clinical measures of pain made at regular postoperative time intervals offer insight into the natural history of pain experienced after knee replacement?

Goodfellow and O'Connor (1986) tried to look at pain trends but this only involved relating the preoperative measure to the latest postoperative measure on each patient.

Do pain measures demonstrate any improvement or worsening over time of the replaced knee?

3. Do pain measures predict later need for reoperation? Does a measure of pain made at the first postoperative follow up have any predictive value? What is the relationship between pain measured in a given postoperative time interval and need of reoperation one or two years hence?

4. Is there any relationship between outcome of two replaced knees within a single patient? If so this has implications for the future reporting of knee replacement series; should the patient or the replaced knee be the unit of observation?

Methods

Patients

Three hundred and sixty four patients were operated upon between June 1976 and December 1989 at the Nuffield Orthopaedic Centre by one surgeon who was closely involved in the design of the prosthesis. Over the period in question, modifications were made in the prosthesis design as experience revealed deficiencies of the original design, especially for certain categories of patient. There are four subtypes of the prosthesis which may be identified in the data. These are

(i) The prototype. This was used in the early years of the series as a test for the bicompartmental model (see (ii)).

(ii) Bicompartmental. This consisted of a prosthesis which replaced both the medial and lateral compartments of the knee. The sliding meniscus which glides in between the femoral and tibial components does not act as a single meniscus

spanning the whole width of the knee but forms two distinct halves for the two compartments. This prosthesis has been less used in recent years of the series.

(iii Medial. This only involves surgery on the inside part of the knee and is useful when the arthritic process has attacked the medial part of the knee alone.

(iv) Lateral. This applies as for medial except that it is used when the outside of the knee has been damaged.

The medial and lateral prostheses have together been responsible for nearly all the recent knee replacements done in Oxford. This decision has been taken on the basis of evolving local experience. Categories (i) and (ii) will be referred to as bicompartmental, and (iii) and (iv) as unicompartmental.

Operation data

Up to four operations were carried out for a given patient. These could consist of a primary operation on either of two knees, and/or a revision operation, and/or a conversion operation. A revision operation is less drastic than a conversion and usually consists of replacing one piece of the prosthesis (the tibial or femoral component or meniscal bearing) which was inserted at the primary operation. A conversion will typically involve removing the prosthesis and replacing it with a more invasive style of prosthesis (eg of Total Condylar design), or even carrying out an arthrodesis or amputation. Thus a conversion betokens an unequivocal failure of the Oxford prosthesis whilst a revision is really an adjustment. Some patients underwent two or even three revision operations of a primary operation done on one knee.

Under such a system there are theoretically an enormous number of permutations of operation types which could be undergone by any single patient, and the distribution of the permutations actually observed in this data set is shown in Table 5-1.

For each operation a large number of items were intended to be recorded. Broadly speaking these relate to decisions taken on details of operative technique, postoperative

management, and observations on the state of the knee.

3) Follow up data

The intended schedule for each patient undergoing an operation was an assessment made preoperatively, six months and twelve months postoperatively and yearly thereafter. The calendar regularity was apt to be altered slightly if a second or subsequent operation took place. The maximum number of "follow up appointments" was 29, although this varied considerably from patient to patient.

At each follow up various assessments were made relating to the patient's general assessment of their knee, the degree of pain experienced while active and resting, ability to walk, climb stairs, rise from a chair and other functional capabilities.

Statistical methods

1. Endpoint definition

Survival curves were constructed where the endpoint was defined in the following alternative ways.

A. The point at which a "conversion" operation was carried out. This involved inserting a relatively invasive total condylar prosthesis and removing the original unconstrained meniscal prosthesis.

B. The point at which a "revision" or "conversion" operation was carried out. A revision operation involved replacing one or more components of the original prosthesis, but leaving the prosthesis itself intact.

C. The date of the clinical assessment when the need for a revision or conversion operation was established. Clearly a long waiting time from the date of this decision until the time of the reoperation would perversely improve the estimated survival time!

D. The date when moderate or severe pain at rest was first noted **or** a decision to reoperate was made.

For those patients who did not reach the "failure" endpoint in the manner specified, their survival time was taken as being from the date of operation to the date of last follow up at which the replaced joint was continuing to function successfully. However there has been suspicion that some groups have published their survival figures assuming that patients not reoperated upon at the time of data analysis were still successes even if they had not been seen for several years (Carr et al 1993a).

The impact on the survival curve of these alternative definitions was investigated. In addition a comparison of the unicompartmental and bicompartmental prostheses was carried out for each defined endpoint using Mantel's version of the logrank test (Mantel 1966).

Survival curves were constructed using the method of Kaplan and Meier (1958). This involves the use of exact follow up times for every patient rather than grouping follow up times into year bands in the manner of Dobbs (1980) and Tew and Waugh (1982). At every follow up time that a single knee prosthesis fails, the survival probability is modified and graphs of survival probability against follow up time show probability falling from 100% downwards in a series of steps . This procedure is demonstrated fully by Peto et al (1977), who provided a formula for the standard error of the survival probability. Thus it is possible to calculate 95% confidence intervals at relevant periods of follow up (eg 5 years and 10 years post operation).

2. Clinical measures and their changes over time

Within each of the follow up periods a four category rating of pain was available. Many other clinical measures were also made but the findings were not particularly different to those found for pain, and so are omitted. The item related to **pain on activity** rather than pain at rest since pain on activity varied between patients to a greater extent.

Pain on activity was assessed as a categorical variable whose categories were "none",

"mild", "moderate" and "severe". The last named was rarely recorded so the categories "moderate" and "severe" were pooled. For a large number of follow ups, the assessment was missing so this was included as an extra category.

The follow up assessments were targeted to take place six months and one year postoperatively and at yearly intervals thereafter. The following postoperative time periods were defined as follows: less than 0.75 years, 0.75 to 1.499 years, 1.5 to 2.499 years, 2.5 to 3.499 years, ... , 6.5 to 7.499 years and 7.5 years or more. Together with the preoperative assessment which was categorised in the same way, this provided up to 10 time groupings. For any patient, no more than one assessment was done in any of these time groupings for the first knee operated upon.

To investigate the relationship between pain measures at successive follow up periods, the rating in the first follow up period was tabulated against the second, the second against the third, the third against the fourth and so on. These tables were then combined so that the nth measure was tabulated against the (n+1)th and any relationship assessed. They were also assessed in terms of whether a consistent trend towards improvement or deterioration could be detected.

The combination of tables involves the repeated use of certain subjects and follows the method described by Bishop et al (1975). This only appears to be justified in terms of **conditional** probability distributions, of a pain measure in the (n+1)th period **given** the pain categorisation in the nth period. It is assumed that the probabilities of change from one category to another is independent of n. Agresti (1990) suggests a more complex modelling approach which expresses the data as a multidimensional table, with the number of dimensions being the number of follow up periods and each dimension of the table including each category of the pain measure. In the present data this would involve 4^9 cells with only 364 patients available so meaningful analysis would clearly be impossible.

In the same way, tables were constructed of the first pain measure against the third, the second against the fourth and so on and then combined into the nth measure against the (n+2)th. This analysis aimed to investigate whether a given measure was predictive of

the same measure two follow up periods ahead and will be referred to as "2 period prediction".

The method was extended so that even a "5 period prediction" was investigated. The prediction was given in terms of positive and negative predictive values concerning the presence or absence of pain a certain number of periods hence.

3. Predictive value of pain for need of reoperation

Pain was again regarded as a four category variable which varied within patients at different follow up periods. To investigate the predictive value of the first postoperative pain measure, a Cox proportional hazards model (Cox 1972) was fitted to estimate the probability of reoperation (using endpoint B: revision or conversion). This analysis will be termed "Early postop".

A second set of analyses was carried out to investigate whether pain measured in a given time interval related to simultaneous need for reoperation. Pain was regarded as a time dependent variable whose value changed as each new follow up interval after operation was reached. The follow up intervals were as outlined in section 2 above. Within each of these intervals one follow up assessment of pain was scheduled, and the time dependent variable assumed throughout the interval the value given at this assessment. This covariate was related to the risk of reoperation during the same interval, again using Cox's proportional hazards model. The analysis will be referred to as "Lag 0".

An analogous analysis investigated the relationship of pain to risk of reoperation in the next follow up interval. At any given time point, the time dependent variable assumed the value of pain in the previous follow up interval (Lag 1). Finally the relationship of pain to risk of reoperation in the next follow up period but one was similarly assessed (Lag 2).

The hazard ratios for reoperation will be presented for all analyses.

4. Use of operations on both knees

In Chapter 2 it was noted that case series are reported with the knee rather than the patient as the unit of observation, and given the substantial number of patients who undergo replacement of both knees, the notion of statistical independence is potentially compromised. The Oxford data carried information of outcome of either knee operation so the following analyses were carried out.

(a) The magnitude of the problem was assessed with regard to the frequency of patients undergoing replacement of both knees. A survival analysis was conducted to assess the length of time between replacing one knee and replacing the second knee. Thus the start point was the first knee replacement and the endpoint the time of the second knee replacement. Patients who only had one knee replaced were regarded as censored at the time of last follow up. Thus it was possible to estimate the proportion of patients receiving replacement of the second knee within a given number of years from the replacement of the first knee.

(b) The length of survival time of the second knee replaced (for those who had both knees replaced) was compared with the length of survival time for patients who had only one knee replaced. This was to assess whether having had the other knee previously replaced had a beneficial effect on the outcome of the second knee replacement.

(c) The possible benefit for the first knee replaced subsequent to the replacement of the second knee (if applicable) was also assessed.

Analysis (a) used the Kaplan Meier technique. Analyses (b) and (c) were conducted using Cox's proportional hazards model. For analysis (b) a fixed covariate (representing whether the patient had one or both knees replaced) was involved. For analysis (c) a time dependent covariate was involved which equalled zero until the second knee was replaced (if applicable) and equalled one subsequently. This type of analysis was used to evaluate the benefits of heart transplantation for patients placed on a waiting list in Stanford, California (Cox and Oakes 1984).

All survival analyses were conducted using the BMDP package (Dixon 1988).

RESULTS

The distribution of operations among the 358 patients is shown in Table 5-1. Sixty three per cent had one knee replaced only with no need for further surgery. A further 23% had both knees replaced with no need for further surgery. Thus only 14% of patients had need of reoperation on at least one of their knees.

1.Definition of outcome

The four definitions of endpoint shown above gave rise to survival curves shown in Figure 5-1. Table 5-2 shows the estimated survival rates at 5 and 10 years post operation respectively.

Clearly since endpoint A (conversion) is a subset of B (revision or conversion), failures are more frequent and survival less for the latter. Only 332 subjects were available for analyses C and D because the date of last follow up was not known for those 24 subjects. Ideally the same number of patients and endpoints should have been present for analysis C as for B (356 patients, 45 of whom underwent revision or conversion surgery). A disproportionate number of "further surgery" patients were lost in analysis C.

Likewise endpoint C (date when need for reoperation established) is a subset of D (date of moderate or severe pain or need of reoperation). According to definition D, twice as many patients reached the endpoint. Hence the alternative definition of failure according to Tew and Waugh (1982) has a considerable yield in terms of numbers of events.

When investigating the relationship between survival and type of prosthesis, for every endpoint unicompartmental prostheses were associated with improved outcome. However differences were more clearly demonstrated for endpoint D as Table 5-3 shows.

For all endpoints the estimated relative risk exceeds 1, confirming the higher risk among bicompartmental prostheses. However according to the confidence intervals, this is not

significant for endpoint A, the endpoint with fewest events. It may also be seen that although the estimated relative risk is similar for endpoints B, C, and D, the confidence intervals are narrower in situations where more events are observed. Thus more information is gained about the relative benefits of the two types of prosthesis where the endpoint is defined such that more events are observed.

2. Clinical measures of pain over time

Relationship between clinical assessments at follow up and subsequent failure

It was clear that a marked shift in pain distribution occurred from the preoperative to the first postoperative assessment. Almost all patients were in considerable pain preoperatively, but almost two thirds were in no pain in the early postoperative period. Because the distribution of preoperative pain was so homogeneous, no further analysis of the preoperative assessment was attempted.

Each follow up assessment was tabulated against the next follow up assessment (1st follow up vs. 2nd follow up, 2nd vs 3rd, 3rd vs 4th etc), and the tables were pooled as shown in Table 5-4.

There was a clear association between degrees of pain on successive occasions. Patients did not appear to be more or less likely to miss their next assessment according to the degree of pain at the current assessment.

One comparison of interest was the direction of change. The table shows that in 127 pairs of assessments the pain became better, while in 119 pairs of assessments it got worse. This would argue against any systematic tendency for pain to increase (or decrease) with time since operation.

This latter analysis was then broken down according to the follow up timing (Table 5-5)

In successive follow up assessments, roughly two thirds of patients remained in the same

pain category at every time point. In the early follow ups (up to four years post operation), changes were more likely to be for the better than for the worse, but in later follow ups the reverse tended to be the case. However the application of McNemar's test for paired samples, did not reveal any significant tendency for change to occur in one particular direction.

The two, three, four and five period prediction tables are shown in Appendix 3. In none of these situations was there any evidence for a consistent systematic tendency for pain to increase or decrease.

Table 5-4 showed that of 70 situations of moderate or severe pain in a given follow up interval, only 34 of these persisted into the next period (positive predictive value of 49%), while of 629 with no or mild pain, 605 continued in this way in the next period (negative predictive value 96%). Table 5-6 shows predictive values of pain assessments for varying numbers of periods hence.

This suggests a gradual weakening of the predictability over time. However at any time point, a lack of marked pain may be expected to continue for at least five periods in over 90% of cases. Experience of pain is more likely than not to resolve itself. Most patients do not experience marked pain at any given time point.

It must be stressed that considerable numbers of observations of pain were absent for these analyses; for example Table 5-4 shows that only 699 of the potential 2912 pairs of observations could be included in the analysis of 1 period predictions. It is plausible that those who continued to attend for **regular** assessments were less likely to alter in their degree of pain than the rest. These observations were all made within the context of an estimated prevalence of marked pain of around 10% at every follow up point.

3. Relationship between pain and later need for reoperation

Early postop

Table 5-7 shows the proportion of reoperations according to pain status at first follow up,

which was not known for 139 out of 358 operations. Of the remainder the likelihood of reoperation increased from 7% for those with no pain to 17% for those with moderate or severe pain. Among those without an assessment, 18% underwent reoperation. The Cox proportional hazards analysis showed no significant difference in hazard between the four categories ($p=0.22$). However there was some suggestion of increasing risk according to early postoperative pain. The fact that those not assessed were at increased risk of reoperation suggests that defaulters from follow up appointments should not be assumed to be representative.

Lag 0

The analysis relating pain to instant risk of reoperation is shown in Table 5-8. Twenty patients whose operation was an instant failure (revision surgery occurred straight after the insertion of the first prosthesis), or who were never seen again subsequent to the operation, were excluded. Pain related strongly to the risk of reoperation, as may be expected since pain is the chief indicator for knee replacement surgery anyway. The greatest risk was for those not attending clinic, but it is likely these patients were not assessed precisely because their reoperation was already planned. The difference between the four categories was highly significant ($p=0.0001$) according to Cox analysis.

Lag 1

Table 5-9 shows the results for relating pain experienced in the period prior to end of follow up or reoperation. Sixty three patients followed up or reoperated upon in less than 0.75 years (9 months) were excluded, together with 20 followed up for more than 8.5 years, leaving 275 patients. Those with moderate or severe pain were more likely to be reoperated upon in the next period (19%) than those with none or mild pain (9%). Those not attending were again at increased risk (16%), possibly suggesting that a missed appointment was predictive of need for reoperation. For some patients the appointment may have been missed because the patient was already on a waiting list for reoperation. Cox modelling revealed a significant difference between the four categories ($p=0.0015$).

Lag 2

Table 5-10 shows the results for relating pain experienced two periods prior to end of follow up to risk of reoperation. One hundred and eight patients followed up for less than 1.5 years were excluded, together with 20 followed up more than 8.5 years. Once again those with moderate or severe pain appeared at greater risk than those with none or mild pain. Those missing an appointment seemed no more likely to undergo reoperation two periods later than those with none or mild pain, a contrasting finding from "Early postop", "Lag 0" and "Lag 1" analyses. However the difference between the four categories was not statistically significant ($p=0.51$).

Table 5-11 summarises the hazard ratios for reoperation of those with moderate or severe pain compared to those with none or mild pain. It may be seen that pain is strongly predictive of need for reoperation either at the time it is assessed or one period later; indeed it is presumably this pain which is largely responsible for the surgeon's decision to reoperate. However there is little firm evidence that pain experienced immediately after the primary operation predicts later need for reoperation, nor does pain clearly predict need for reoperation two periods later.

4. Use of operations on both knees

(a) Figure 5-2 shows the likelihood of a second knee replacement within a given period from the first knee replacement. Out of 347 patients with known dates on all relevant operations, 98 underwent a second replacement of whom 36 had both knees replaced within the same operation (10%). The survival curve shows 22% would be predicted to have undergone a second operation within one year and 33% by five years. By the maximum length of follow up (12.5 years), this figure has increased to 36%. Thus most second operations were done within a year of the first.

Data pertaining to analyses (b) and (c) are shown in Table 5-12. It may be seen that of 260 patients for whom the second knee was not replaced, 35 of the "first" knees failed (13%), whereas 10 of the first knees failed out of the 98 whose second knee was replaced (10%). The difference is reversed when the 36 patients who had both knees replaced

simultaneously are removed from analysis. The failure rate of first knees now becomes 10 out of 62 (16%), since none of the 36 patients experienced failure of either knee replacement.

(b) Figure 5-3 shows a comparison between the survival of 2nd knees replaced with unilaterally replaced knees, until the date of a revision or conversion operation. Of the 358 patients shown in Table 5-12a, only 343 patients had complete data available for survival analysis. Six out of 91 second knees (7%) failed as opposed to 35 out of 252 unilaterally replaced knees (14%). The former were followed up for less time but no failures were seen within two years of operation. The relative hazard was 2.38 for patients with only one replacement (95% confidence interval 1.0 to 5.65; p=0.05 by Cox's regression analysis). When the 36 patients who had both knees replaced simultaneously were excluded, the relative hazard was 2.21 (95% confidence interval 0.86 to 5.66, p=0.10). The weak evidence of a difference may be due to selection for a second knee replacement being conditional on promising results with the first.

(c) Figure 5-4 compares the unilaterally replaced knees with the survival of the first knee in bilaterally replaced patients. The curve for the latter group does not show at what point the second knee was replaced. Ten out of 92 first knees failed (11%). However Cox's model suggested that the relative hazard for patients not undergoing a second operation was 1.69 (95% confidence interval 0.84 to 3.46; p=0.14 from Cox's model). Thus there is no significant evidence for a protective effect on the first knee when the second knee is replaced. Exclusion of patients with simultaneous bilateral replacement gives a relative hazard of 1.37 (95% confidence interval 0.67 to 2.79, p=0.38)

From Table 5-12 the crude failure rate for all first knee operations is 45/358 (13%) and for all knees combined it is 51/456 (11%). A more refined method which estimates probabilities of 5 year survival from data in Figures 5-3 and 5-4 suggests the probability of failure at 5 years would be 16.4% for first knee operations and 14.8% for all knee operations. The difference would be amplified if the series contained a higher proportion of patients with both knees replaced (see Appendix 4).

Finally Table 5-12 shows no evidence of association among bilaterally replaced patients

between results for the two knees, since none of these 98 patients experienced failure of both knee replacements.

Discussion

Alternative definitions of endpoints for survival analysis led to dramatically different results, with an almost three fold difference in failure rates reported. Where case series are reported in the orthopaedic literature with no standardisation of outcome definition (see Chapter 3), widely differing reported success rates mean little. Moreover although differing failure rates between two subgroups tend to go in the same direction whichever endpoint definition is chosen, differences are more precisely estimated when the endpoint leading to the greatest number of events is chosen. This has major implications for comparative research.

Scuderi et al (1989) report a series of 1430 operations involving a comparison of three subtypes of the same prosthesis. However the total number of events (revision operation carried out or scheduled) was only **25**. Whilst this doubtless demonstrates excellent technical success it allows little chance of demonstrating differences between prosthesis subtypes, even in such an enormous series. Following Tew and Waugh (1982) a number of authors have reported survival not only in terms of the prosthesis remaining **in situ** but also in terms of the knee being free of severe pain. Thus Thacker and Fulford (1986) in a series of 335 knee replacements reported 9 events according to the first criterion and 21 according to the second.

Multicentre surveys have been able to demonstrate differences between subgroups very effectively; Knutson et al (1986) reported 649 failures among 8000 arthroplasties followed up to eight years while Rand and Ilstrup (1991) showed 1107 failures among 9200 arthroplasties followed between two and 18 years. From these authors' experience it is likely that when failure rates are below 10%, thousands of knee replacements need to be studied in order to detect differences between subgroups of patients treated with different prostheses. Given the high survival rates reported after 10 to 15 years, it would also be necessary to follow them up for a long period in order to observe more events.

The need for a more sensitive outcome observable after a relatively short period is apparent. The analysis of the present data set showed that while existence of marked pain is more common than revision surgery, pain does not seem to be a very predictable entity. Over half the patients reported as suffering marked knee pain at a given follow up appointment were better at the next follow up. In addition a single measure of pain at an early postoperative appointment did not appear to be very useful in predicting the later need for revision surgery. Need for reoperation was only related to pain experienced less than two years previously. However pain may still be a helpful measure in comparing large groups of patients. Its lack of long term relationship to need for reoperation does not invalidate the measure. The measurement of pain from a postal questionnaire rather than from outpatient assessment is reported later in this thesis.

All reported series have regarded the knee rather than the patient as the unit of observation in reporting success, thus double counting patients with both knees replaced. The analysis above suggests such an approach may have led to moderate bias in the direction of overoptimism, since bilaterally replaced patients appear to have experienced a lower failure rate than unilaterally replaced patients. In comparing results reported in the literature, account needs to be taken of the proportion of patients who underwent bilateral replacement.

A different issue concerns the precision of estimates of survival. Confidence limits for survival rates have been promoted recently in the orthopaedic literature by Lettin et al (1991) and Nellisen et al (1992). This study has been unable to demonstrate any significant intra patient correlation in outcome of replacement of the two knee joints. We may not assume that no such correlation exists, but the effect can only be estimated with a series in which at least 1000 patients have undergone bilateral replacement. The effect may turn out to be small enough to ignore, according to the criterion of Rosner (1982).

The implications for comparison of survival between two treatment groups are interesting. Firstly, whether patients are unilaterally or bilaterally replaced should be used as a stratifying factor, since survival seems better for the latter group. However this is not enough to compensate for possible interdependency in results for bilateral replacements.Both Rosner (1982) and Donner and Donald (1988) have shown that if such

dependency is ignored, significance levels will be exaggerated. Both have suggested ingenious methods to correct for the effects of such dependency when the outcome is binary. Unfortunately this is inadequate for subgroups of patients who may have been followed for differing lengths of time. A still more difficult problem arises when surgeons compare their results between knees replaced with prosthesis Mark I with knees replaced with Mark II, since some patients will have Mark I implanted in one knee and Mark II implanted in the other. Such data are liable to be riddled with selection biases and should be discouraged. Moreover the methods of Rosner, and Donner and Donald, assume that multiple observations made on a single patient may be attributed to the same treatment.

It may be possible to extend Donner and Donald's method by stratifying survival data into years of follow up and combining the strata in the manner of Mantel (1966). However the dependency effect is likely to be small, and the method rather cumbersome, so that little would be gained. In a randomised trial comparing survival for patients treated with different prostheses, very little difference in significance levels would be seen between using only one joint per subject and using both joints in bilaterally replaced patients as if they were independent. An extended version of Donner and Donald's method would presumably give an intermediate significance level.

A related issue not explored here concerns the inclusion of primary and revision operations on the same subjects. Rand and Ilstrup (1991) included both primary and revision operations in their sample of 9200. These authors demonstrated a markedly lower survival rate among revision operations than primary operations. Among certain patient subgroups at higher risk of failure (for example those with osteoarthritis), a certain amount of double counting would occur, thus overestimating the relative risk of failure. This dubious practice of counting both primary and revision operations in the same analysis should be discouraged.

Conclusions

Reported survival depends crucially on the definition of the endpoint. In the modern era conventional endpoints such as revision surgery are unlikely to occur with sufficient

frequency to detect small benefits of one prosthesis over another. Pain assessed at follow up appointments varies considerably within patients and does not promise to be a useful long term predictor of need for reoperation.

The use of operations on both knees of particular patients as if they were independent is intuitively suspect. However this study suggests that this is unlikely to lead to gross errors of estimating survival probabilities for one particular knee operation. A very large series would be required to elucidate whether the independence assumption is valid.

Table 5-1. Distribution of the permutations of operation types among the 358 patients

Description of operations experienced	No. of patients
One knee replaced, no further operations	225
Both knees replaced, no further operations	82
One knee replaced, revised once	10
One knee replaced, later converted	13
Both knees replaced, one revised once	8
Both knees replaced, one converted	5
One knee replaced, revised twice	3
One knee replaced, revised once, then converted	7
One knee replaced, revised three times	1
One knee replaced, revised twice, then converted	1
Both knees replaced, one revised once and later converted	3
Total	358

Table 5-2. Survival rates at 5 and 10 years postoperation according to four different definitions of endpoint.

End-point	No. patients:		5 years		10 years	
	In analysis	Reached endpoint	No. at risk	% (se) survival	No. at risk	% (se) survival
A.	356	26	112	90(3)	16	79(9)
B.	356	45	102	84(3)	13	72(11)
C.	332	36	97	86(3)	12	78(11)
D.	332	74	74	70(4)	9	59(13)

Table 5-3. Comparison of numbers of events between patients treated with bicompartmental and unicompartmental prostheses, according to four different endpoints.

Endpoint	Bicompartmental: no. of events		Unicompartmental: no. of events		Relative risk (95% CI)
	Observed	Expected	Observed	Expected	
A	22	20.13	4	5.87	1.68 (0.56-5.06)
B	40	32.12	5	12.88	3.42 (1.32-8.86)
C	34	26.61	4	11.39	3.89 (1.36-11.1)
D	65	49.64	9	24.36	3.73 (1.84-7.55)

Table 5-4. Relationship between degree of pain experienced at successive follow up appointments

Pain during activity

(n+1) th follow up

nth follow up

	None	Mild	Mod/Sev	Missing	Total
None	352	95	8	300	755
	(47%)	(13%)	(1%)	(40%)	(100%)
Mild	91	67	16	123	297
	(31%)	(23%)	(5%)	(41%)	(100%)
Mod/Sev	17	19	34	47	117
	(15%)	(16%)	(29%)	(40%)	(100%)
Missing	184	66	33	1460	1743
	(11%)	(4%)	(2%)	(84%)	(100%)
Total	644	247	91	1930	2912

(In this table, missed follow ups were included with the missing category)

The unit of observation is not the patient but the paired consecutive assessment on any patient. Thus an individual patient may contribute several observations to this table.

Table 5-5. Changes in degree of pain during activity in various successive follow up appointments

Follow up:	Result: Better	Same	Worse	Total
1st to 2nd	29 (13%)	167 (77%)	20 (9%)	216 (100%)
2nd to 3rd	33 (26%)	65 (52%)	27 (22%)	125 (100%)
3rd to 4th	23 (17%)	97 (70%)	18 (13%)	138 (100%)
4th to 5th	18 (17%)	68 (65%)	19 (18%)	105 (100%)
5th to 6th	10 (13%)	56 (71%)	13 (16%)	79 (100%)
6th to 7th	8 (13%)	41 (64%)	15 (23%)	64 (100%)
7th to 8th	9 (21%)	27 (64%)	6 (14%)	42 (100%)
8th to 9th	3 (13%)	16 (67%)	5 (21%)	24 (100%)

(Observations on patients where at least one of the two follow up assessments is missing, are omitted)

Table 5-6. Positive predictive (PPV) and negative predictive values (NPV) for existence of moderate or severe pain at varying numbers of time periods hence

Number of periods	PPV	NPV
1	34/70 (49%)	605/629 (96%)
2	22/58 (38%)	481/514 (94%)
3	19/46 (41%)	356/376 (95%)
4	9/26 (35%)	245/263 (93%)
5	6/20 (30%)	151/165 (92%)

Table 5-7. Relationship between pain at first postoperative follow up and subsequent reoperation. Numbers (percentages shown)

Pain	Reoperation		
	No	Yes	All
None	126(93)	10(7)	136(100)
Mild	49(91)	5(9)	54(100)
Moderate /severe	24(83)	5(17)	29(100)
Not done	114(82)	25(18)	139*(100)
Total	313(87)	45(13)	358(100)

* Of these, 2 attended the clinic but pain was not assessed

Table 5-8. Relationship between pain at the end of follow up and immediate risk of reoperation (Lag 0). Numbers (percentages shown)

Pain	Reoperation		
	No	Yes	All
None	178(95)	10(5)	188(100)
Mild	64(93)	5(7)	69(100)
Moderate /severe	22(73)	8(27)	30(100)
Not done (but patient attended clinic)	19(95)	1(5)	20(100)
Not done (patient did not attend)	12(39)	19(61)	31(100)
Total	295(87)	43(13)	338(100)

Table 5-9. Relationship between pain one period prior to end of follow up and risk of reoperation (Lag 1). Numbers (percentages shown)

Pain	Reoperation		
	No	Yes	All
None	99(89)	12(11)	111(100)
Mild	37(97)	1(3)	38(100)
Moderate /severe	16(76)	5(24)	21(100)
Not done	88(84)	17(16)	105*(100)
Total	240(87)	35(13)	275(100)

* Of these, 5 attended clinic but pain was not assessed

Table 5-10. Relationship between pain two periods prior to end of follow up and risk of reoperation (Lag 2). Numbers (percentages shown)

Pain	Reoperation		
	No	Yes	All
None	96(89)	12(11)	108(100)
Mild	46(87)	7(13)	53(100)
Moderate /severe	8(73)	3(27)	11(100)
Not done	52(90)	6(10)	58*(100)
Total	202(88)	28(12)	230(100)

* Of these, 2 attended clinic but pain was not assessed

Table 5-11. Hazard ratios for reoperation with regard to pain assessed as moderate or severe compared with none or mild pain

Time of pain assessment	Hazard ratio	95% confidence interval
Early postop	2.04	0.74 to 5.62
End of follow up (Lag 0)	4.70	2.00 to 11.0
One period before end of follow up (Lag 1)	4.19	1.48 to 11.9
Two periods before end of follow up (Lag 2)	1.79	0.52 to 6.19

Table 5-12a. Relationship between outcome of first and second knees replaced (including 36 simultaneous bilateral replacements)

Outcome for 2nd knee replaced

Outcome for 1st knee replaced	Not replaced	Successful at last follow up	Failed	Total
Successful at last follow up	225	82	6	313
Failed before 2nd knee replaced	35	2	0	37
Failed after 2nd knee replaced	--	8	0	8
Total	260	92	6	358

Table 5-12b. Relationship between outcome of first and second knees replaced (excluding 36 simultaneous bilateral replacements)

Outcome for 2nd knee replaced

Outcome for 1st knee replaced	Not replaced	Successful at last follow up	Failed	Total
Successful at last follow up	225	46	6	277
Failed before 2nd knee replaced	35	2	0	37
Failed after 2nd knee replaced	--	8	0	8
Total	260	56	6	322

(Failure = Revision and/or conversion operation carried out)

Figure 5-1. Survival of Oxford knee replacement series according to different endpoints

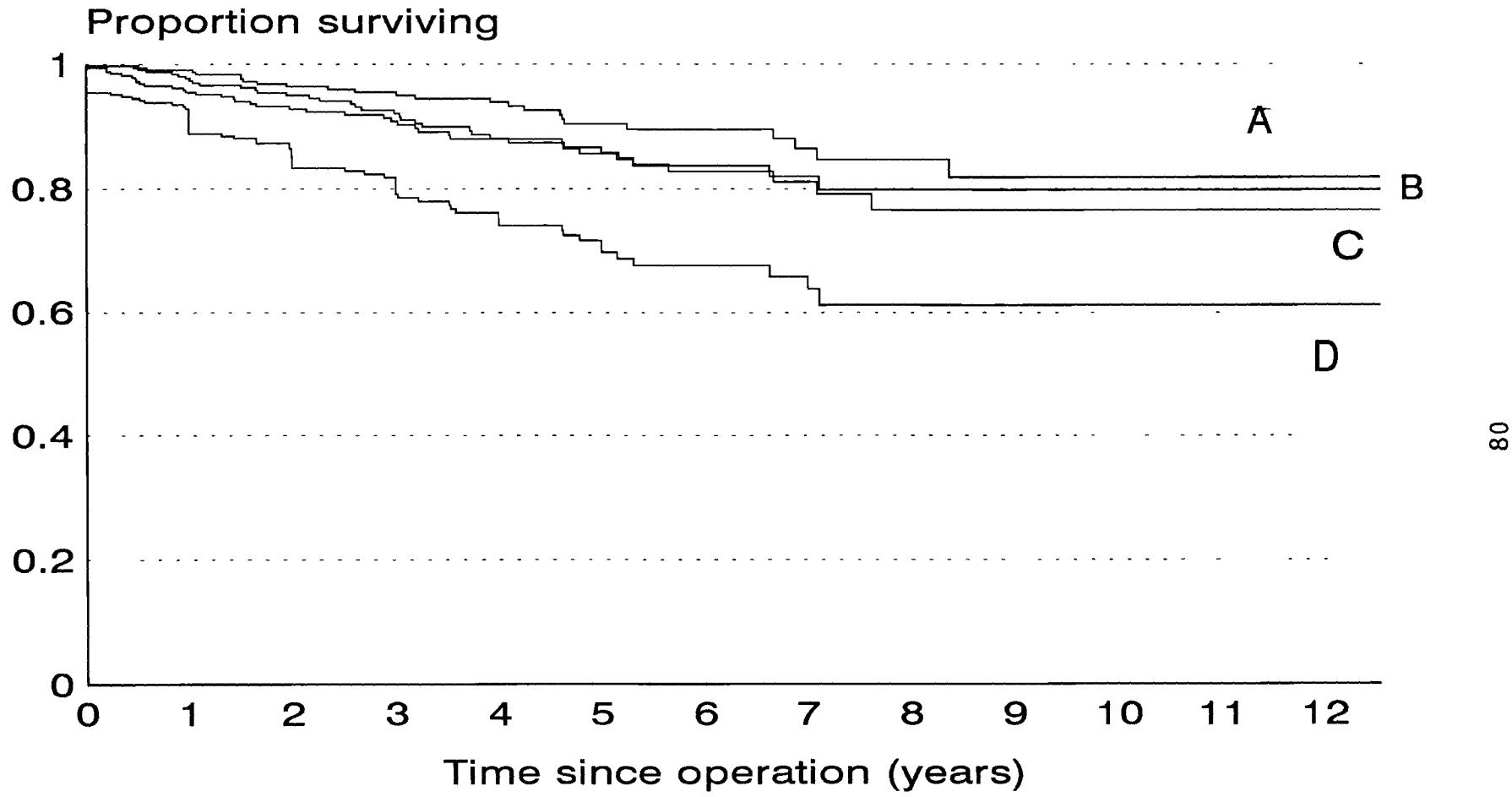


Figure 5-2. Survival curve to show time between first and second knees replaced in Oxford series

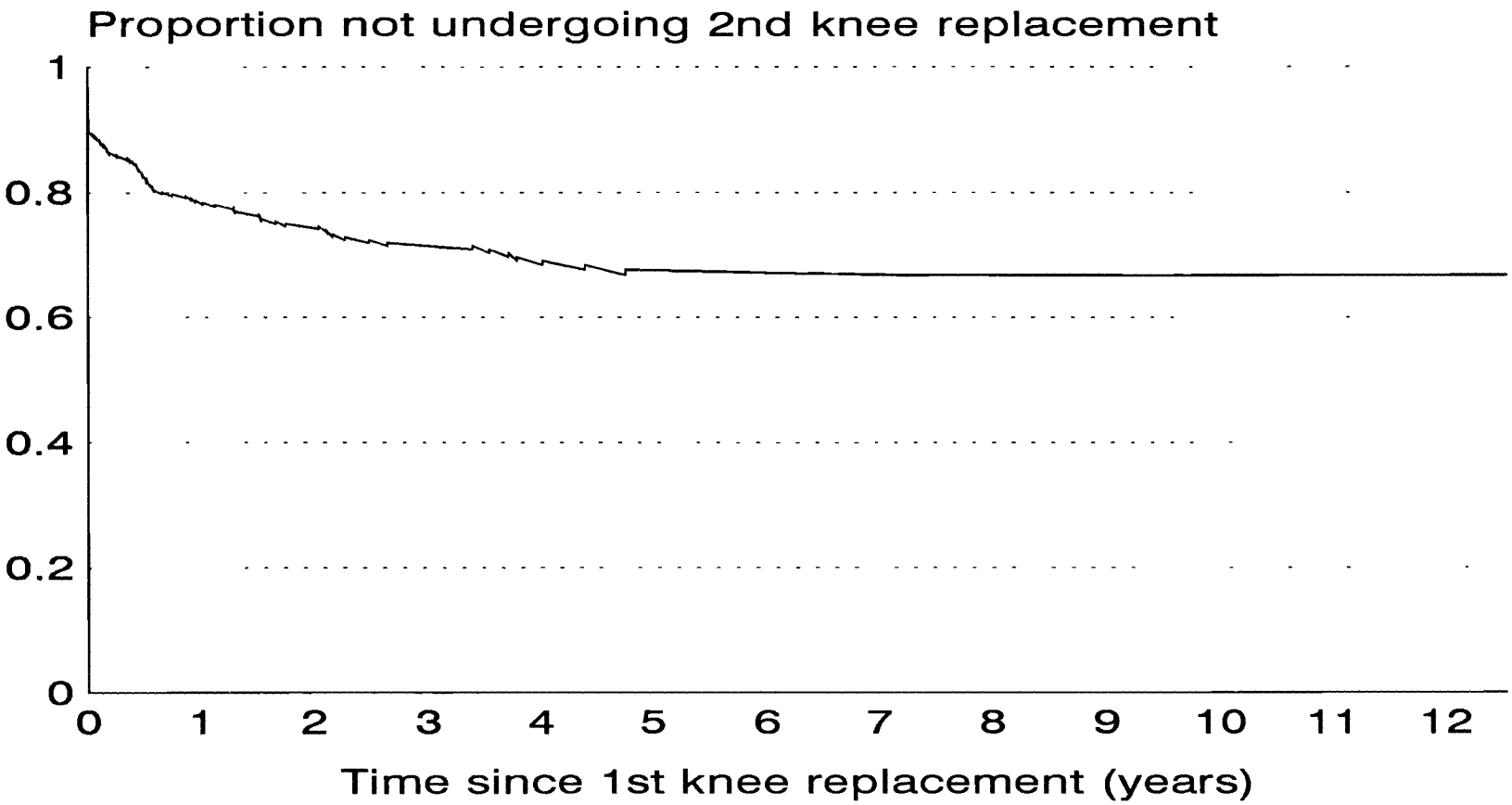


Figure 5-3. Comparison of survival between patients undergoing replacement of one knee vs those undergoing a second knee replacement

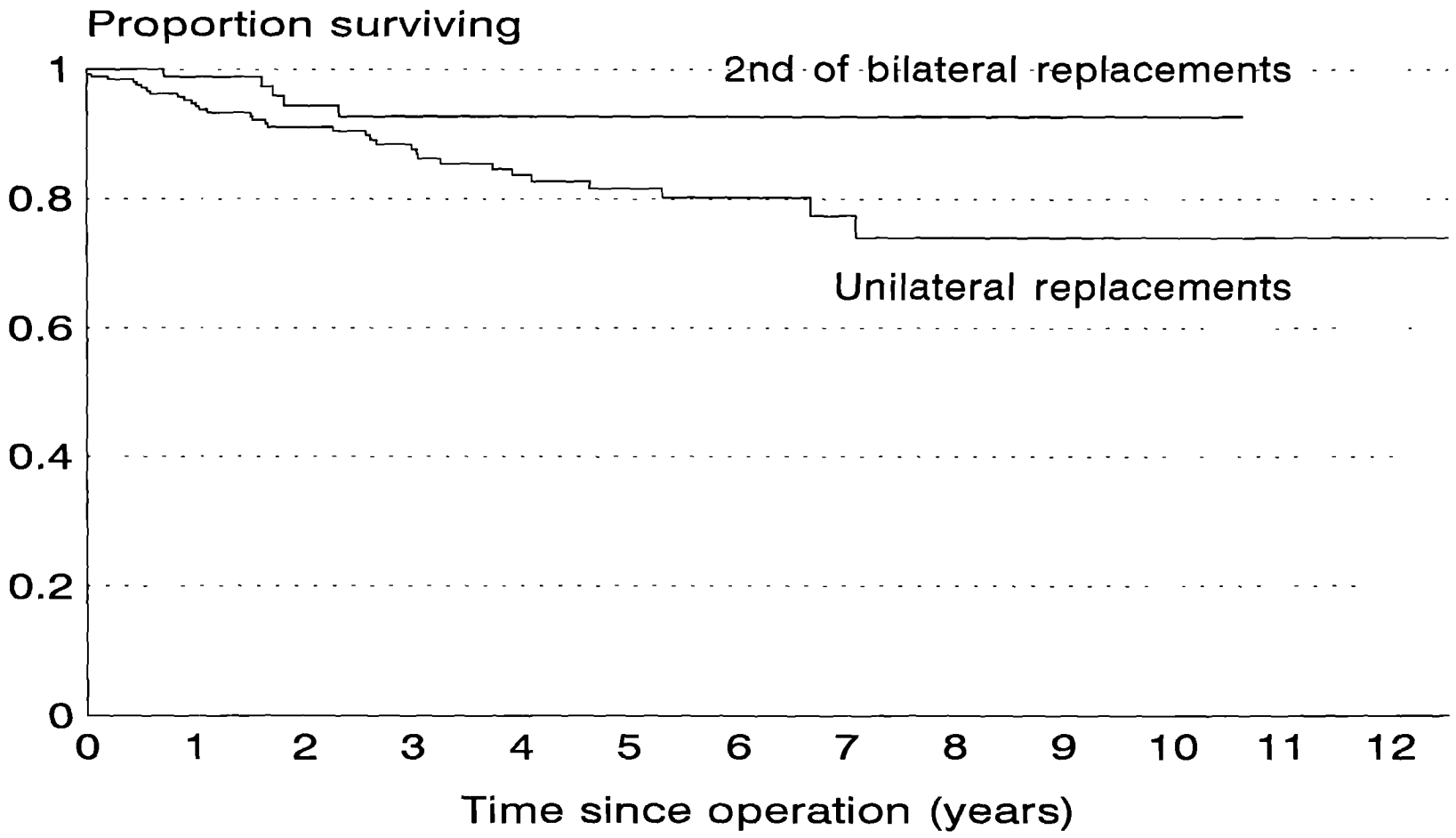
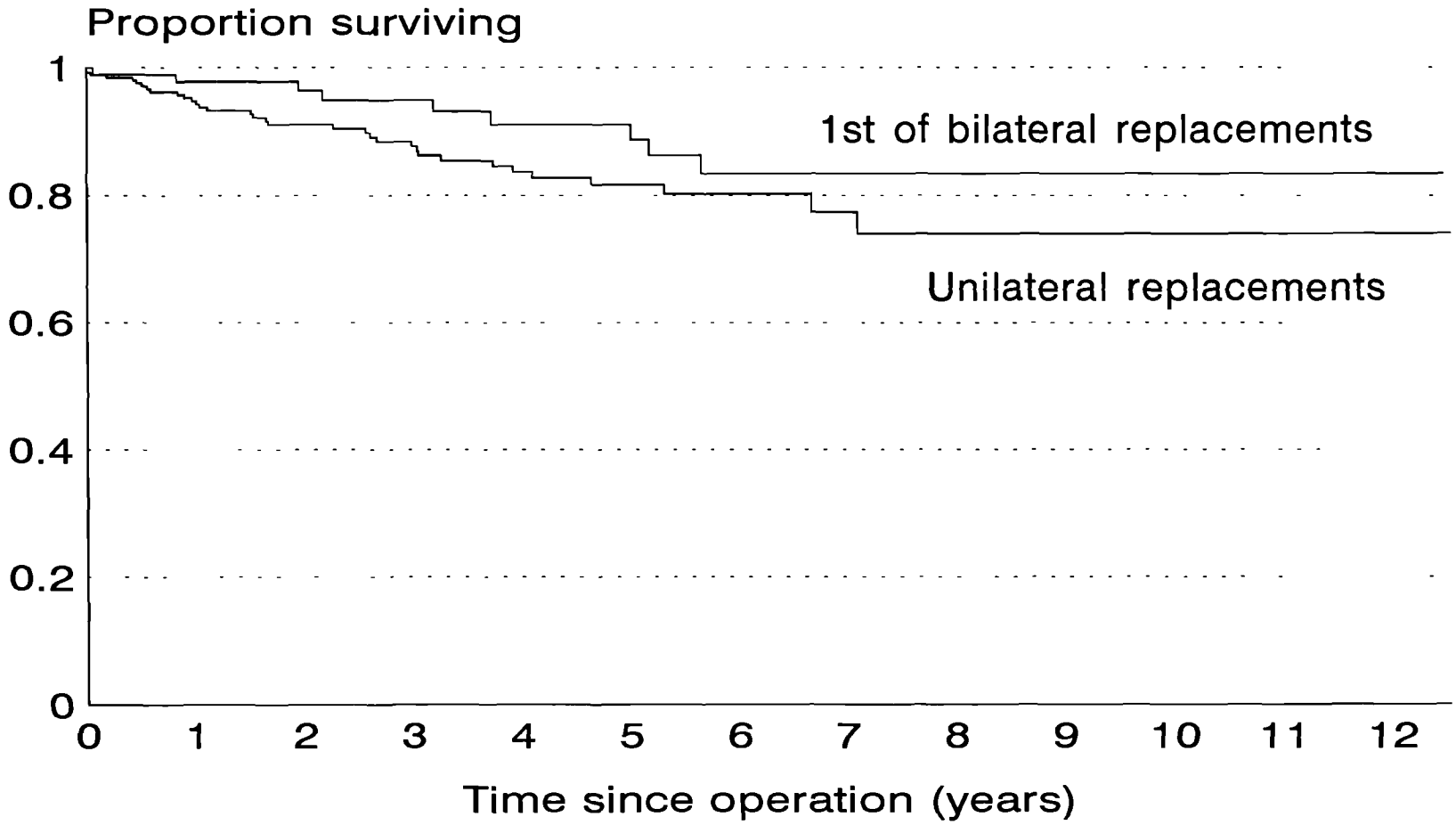


Figure 5-4. Comparison of survival for patients undergoing replacement of one knee vs those undergoing first of two knee replacements



CHAPTER 6. SELECTION OF SERIES FOR QUESTIONNAIRE SURVEY

Initially the intention was to survey four series of patients, each of whom had been under the care of one particular surgeon using one particular prosthesis. These four surgeons are in UK terms pioneers of knee replacement operations and had each begun their series of patients in the 1970s. Each had at the time of this survey carried out operations on more than 300 patients. Each has in his own way been enterprising and creative in searching for optimal treatment for patients suffering disabling and painful arthritis of the knee joint.

It was realised however that a comparison of results between the four groups of patients was not simple. Even if the four groups came from similar populations (with regard to such aspects as gender, age, weight and disease severity), and even if a standardised outcome variable could be used, any difference in outcome detected could not be imputed automatically to the prosthesis. Knee replacement demands a large measure of surgical skill. Any difference seen in outcome between the four groups of patients may be due to the prosthesis, due to the surgeon, or a combination of the two factors. Some authors have drawn attention to the surgeon effect in evaluating alternative operations (Dudley 1985, Pollock 1989). Additionally postoperative rehabilitation policy may well vary between units.

Hence further series of patients were added to gain insight into the size of the surgeon effect.

Secondly the populations of patients under the care of different knee surgeons may differ according to the type of hospital they operate in, and the social profile of the catchment area.

The series eventually included in the study were composed as follows:

01- Inner London teaching hospital. Prosthesis A, not invented by surgeon (S1)

02- Inner London non teaching hospital. Prosthesis A, also not invented by this surgeon(S2)

03- Inner London teaching hospital. Prosthesis B, invented by surgeon, not used elsewhere.(S3)

04- Inner London teaching hospital. Prosthesis C, invented by surgeon, not used widely elsewhere (S4).

05- Provincial district hospital. Prosthesis D, not invented by this surgeon (S5) but invented by surgeon S8(see below)

06- Inner London specialised orthopaedic hospital, attached to teaching hospital. Prosthesis C, Surgeon S4 (see above for series 04)

07- Provincial specialised orthopaedic hospital. Prosthesis E, invented by surgeon (S7). Used in a number of other centres.

08- Provincial district hospital. Prosthesis D (see above for Series 05), invented by surgeon (S8).

09- Inner London teaching hospital. Prosthesis F, invented by surgeon (S9). Used extensively in other centres, not only in UK.

10- Provincial district hospital. Prosthesis E, not invented by surgeon (S10), but by surgeon S7 (see above)

The relationship between the surgeons, the prostheses they used and the series number by which they will be referred are shown in Table 6-1.

Thus the data on Prostheses A,D,E gave information from which the surgeon effect could be investigated. Outcome in patients treated by the prosthesis inventor may be better than

for patients treated by another surgeon with that same prosthesis, and data for Prostheses D and E are useful to investigate this effect. Prosthesis A was invented by neither surgeon S1 nor S2, but S1 was more experienced in knee replacement generally and began using prosthesis A before S2.

It should be noted that prosthesis E was first designed to replace both sides, or compartments of the knee joint but in recent years it has been more commonly used as a unicompartmental design in which only the side of the joint affected by arthritis is replaced (usually the medial compartment but occasionally the lateral). Series 07 includes a mixture of patients treated with the unicompartmental and bicompartmental designs but Series 10 includes only the bicompartmental treatment. Analysis in Chapter 10 will distinguish between these two designs.

Data for prosthesis C would allow the investigation of differences between results obtained by a single surgeon at two different hospitals. However Series 04 and 06 are distinguished chiefly because different methods were used for lifting names and addresses of the patient groups.

Sampling procedures

Series 01

Surgeon S1 had kept careful records of operations carried out with prosthesis A since 1983. It must be stressed that this surgeon sometimes used other prostheses, but prosthesis A was the treatment of choice. Surgeon S1 kept special records only of patients operated upon with prosthesis A.

As for several of the other patient series, a list of the hospital numbers for each patient was the means by which patient names and addresses were elicited, from the hospital's computerised records system. This system also had some (though incomplete) information on whether patients had died.

Surgeon S1 attempts to carry out regular outpatient follow up of these patients. At the time of the study these happened no more than three times a year. Thus follow up was not sufficiently complete to allow comparison of clinical outcome and patient responses to the postal survey.

Series 02

Surgeon S2 had begun a series of patients with prosthesis A soon after Surgeon S1. The same means was used for tracing patient names and addresses.

Surgeon S2 also attempted to keep regular follow up records but these were even less complete than for Surgeon S1.

Series 03

Surgeon S3 had been operating with Prosthesis B since 1971 at different centres within his district, and estimates that he had carried out around 740 operations prior to 1990. Unfortunately one of his main operating centres had recently been closed and hospital records had been incinerated. Over a six month period Surgeon S3 agreed to list patients' names as they arrived spontaneously at outpatient clinics and thus a subset of his patients were accrued. There are no means of checking whether the 48 patients accrued were representative of the whole.

Series 04

Surgeon S4 kept records of all knee replacements carried out from 1973 onwards until 1982. From 1982 all his knee replacements were entered into a multi centre study of knee replacements (EULAR). The coordinating centre for the EULAR study were able to provide a printout of hospital numbers, and these were also available for the records before 1982.

For Series 06, the same applied.

The head of the patient records department at the hospital provided patient names and addresses, taken from the hospital computer.

Series 05

Surgeon S5 had been operating at a number of district general hospitals in his area since 1981. In 1990 a research registrar carried out a review of the series and compiled a list of patients simply by scanning theatre operating records from the various hospitals. From this a list of 146 patients was compiled. The research registrar had already compiled a list of addresses of this group.

Series 06

The same sampling procedure was carried out as for Series 04. RWM went to the hospital records department and extracted names and addresses off the hospital computer and from filed patient notes.

Series 07

Surgeon S7's collaborator in long term follow up of the operated patients keeps a file of patient data, with names and addresses included. Addresses turned out to be available for 239 patients. Of the remaining 125, all but 36 could be found from the hospital records. Of the remainder, some turned out to be private patients and their addresses were identified by surgeon S7's private secretary.

Surgeon S7 has been meticulous in his recording of patients operated upon with prosthesis D, and rarely used any other prosthesis for a few years after his series began in 1976. In recent years he has tightened his selection criteria after analysis suggested good outcome was more likely for OA patients where the cruciate ligaments were intact.

Series 08

Surgeon S8 has also carried out a meticulous follow up of his patients, for whom the series also began in 1976. Surgeon S8 keeps all patient data on his own computer. A copy of names and addresses was lifted directly from this.

Surgeon S8 has now retired from NHS work, and his list of patients does not include the patients operated upon privately either before or since his retirement. This surgeon has also carried out a number of operations for teaching purposes at other hospitals, and these are also not included in the present series.
The patients included for this study comprise approximately 90% of all patients operated upon by Surgeon S8.

Series 09

Surgeon S9 has probably the largest and longest series of knee replacement operations in this country. He has kept records in a special office which employs two part time staff to maintain the records and organise regular follow up. In May 1990, names of 543 patients were found on the his data base. It was considered expedient to take a one in two random sample resulting in 262 patients, using random number tables. Unfortunately the data base contained neither hospital numbers nor addresses. Manual records revealed hospital numbers for 172 of these and special permission was granted to access the hospital computer for up to date addresses. This was successful for 152 patients, of whom 7 turned out to be duplicates.

Ninety of the patients were not on this hospital computer because they were private patients. The time required to access and use the computer system held at Surgeon S9's private clinic to obtain a list of the private patients' addresses was not considered to be within the resources available for this study.

Hence in all a sample of 145 patients was obtained, all NHS. This may be biased since most of the other series contained a small proportion of private patients. It was considered however that inclusion of surgeon S9's private patients, which comprised a larger proportion of his total patient population than was the case for other series, would

have biased his series towards a relatively privileged patient population.

Series 10

Operations carried out at this hospital were not only carried out by the consultant in question but on perhaps one third of occasions by his senior registrar. It was in fact the current Senior Registrar who helped RWM to discover a batch of forms which was purportedly a prospectively formed list of patients operated upon. Seventy six patients were found in this list, compared with 62 reported by Barrett et al (1990) for this series. The discrepancy is explained by a few patients operated upon subsequent to the time when Barrett et al carried out their analysis.

Effect of sampling procedure on comparison of series

Nearly all this series consisted of patient groups formed prospectively by the surgeons. The two exceptions were Series 03, which is likely to comprise a biased sample, and Series 05 which was collected by retrospectively inspecting theatre records. Surgeon S5 felt he had carried out around 200 operations and 146 patients were actually identified. The respondents in Series 05 in the questionnaire survey suggested one third of patients had both knees replaced, so it is probable that the 146 patients comprised the great majority of those actually operated upon.

For all surgeons except Series 03 and Series 09, very few other operations seem to have been carried out other than those identified in this study. The exclusion of private patients for Series 09 would only be important if private patients differ markedly from the rest in their outcome. The other series did not have sufficient patients to explore this.

The actual profiles of the patient series, and the response rates obtained to the questionnaire, will be explored in Chapter 8.

Table 6-1. Relationship between surgeons, series and prostheses.

Prosthesis	Series	Surgeons	Type of prosthesis
A	01,02	S1,S2	Minimally constrained
B	03	S3	Minimally constrained
C	04,06	S4	Constrained Hinge
D	05,08	S5,S8	Semi constrained
E	07,10	S7,S10	Compartmental, unconstrained
F	09	S9	Minimally constrained

CHAPTER 7. QUESTIONNAIRE DEVELOPMENT AND MAILING

Since an aim of this thesis was to develop a patient centred measure of outcome, I sat in on a number of knee replacement follow up clinics held by two different surgeons. In one clinic I was permitted to ask patients questions concerning their experience with their knee replacement, to gain some understanding of the potential problems.

Many general health questionnaires have been developed to measure day to day wellbeing such as the Sickness Impact Profile (Bergner et al 1981) and the well used Nottingham Health Profile (Hunt et al 1984). The Nottingham instrument is really far too general for the purpose in hand and relates more to the impact of disease. The same may be said for the SIP, although a British equivalent was designed for a study of disability in the community (Patrick 1981). Both British and American versions of the SIP contain 136 items. A promising alternative was the Stanford Health Assessment Questionnaire (HAQ) which has been used for rheumatoid arthritis patients (Fries 1983). In particular it is currently in use for a longitudinal study of patients entered on to the waiting list for joint replacement (hip or knee) at the London Hospital (Kirwan et al 1992). However it is once again designed to get a more general picture of the range of restrictions in daily activities experienced by arthritis sufferers. Some of the questions relate directly to arthritis in upper limb joints (eg ease of cutting up food, ability to raise hands above head), and several are intended for American subjects. Much the same may be said for the Arthritis Impact Measurement Scale (AIMS: Meenan et al 1980) which has been carefully validated in arthritis sufferers. It comprises 45 items and takes 15 minutes for a patient to complete.

Mention should be made of the WOMAC scale, designed in Canada to assess patients with osteoarthritis of the hip and knee joint (Bellamy and Campbell 1989). An attempt to validate it was made on a sample of 30 patients who had undergone hip or knee replacement (Bellamy et al 1988). The scale comprises 42 items, of which 30 changed significantly with surgery. It would appear the most useful published index for the type of research carried out in this thesis. Nonetheless with 42 items on five different dimensions, it is still quite long. It has not been used for a British population and a review of general outcome measures for orthopaedics made no mention of it (Radford

1993).

Given that the aims of this thesis were to assess success of the knee replacement itself, a special questionnaire was felt necessary. Because arthritis strikes at multiple joints and because it occurs chiefly among the elderly, questions whose response may be imputed solely to troubles with the replaced knee joint are difficult to design. The questionnaire eventually used had the following thrusts:-

(i) Range of activities required in daily living

(ii) Questions on walking aids used, joints which have been replaced, and joints which give pain.

(iii) Direct questions on patients' opinions of the success of the knee replacement operation, and recent pain experienced.

(iv) Sociodemographic variables

Activities are not restricted just by a single joint, although activities were selected which may well be related to the knee joint. For example toe nail cutting is something which is made difficult by arthritis of the hip as well as the knee. In addition certain activities may be eased by a knee replacement, but coexisting arthritis of the opposite knee may still cause problems. For this reason it was decided to collect data on evidence of arthritis in other joints, either those which have been replaced or those recently giving pain. These could be confounders in any analysis which compares results between different prostheses.

Among other confounders may be social variables. In particular the amount of support obtained from relatives and the suitability of accommodation may determine a patient's recovery from an operation and subsequent long term success of their artificial joint. Volumes of research have gone into the social class gradient of levels of health in the community (Townsend and Davidson 1982) and this could well include recovery from a

knee replacement operation. Data on the patients' social class according to the Registrar General's Classification of Occupations may not be appropriate for a largely retired population. Marked differences in Standardised Mortality Ratios according to housing status have been found in the OPCS Longitudinal Study (Fox and Goldblatt 1982). This appeared a good social variable to include in the analysis because housing unsuitability may per se cause difficulties in mobility and experience of success with the knee replacement. Lastly having a relative living with the patient may alleviate some of the agonies of daily activities and promote positive attitudes towards the knee replacement. This does not mean that those living alone are necessarily expected to do worse since they may be a selectively fit group, whilst those needing a relative or friend to stay may well be more severely disabled.

Direct questions on patients' opinions were included since these could be compared and contrasted with clinicians' opinions of outcome. These were designed so that answers were unlikely to be influenced by pain in other joints.

Sources of questions asked

(i) Functional items.

These were partly taken and modified from the Stanford HAQ (Kirwan and Reeback 1986) and partly related to items seen in the BASK questionnaire (Aichroth et al 1978). They also reflect priorities seen in most knee assessment forms (Insall et al 1976, 1989). Those activities mentioned by patients in the knee follow up clinics I attended were also considered.

(ii) Patients' opinions, pain experience

Several visual analogue scales were included with crosses marked to define a score from 0 to 6 (see Appendix 5). Discussion with a pain clinic consultant resulted in the wording chosen. A sheet was included for left and right knees separately to allow for patients who had undergone replacement operations of both knees. Of the seven visual analogue scales

included the first five were constructed such that 0 indicated a bad result and 6 a good result. This was reversed for the last two visual analogue lines which both related to pain experience.

At the end of the section of knee oriented questions were two multiple choice questions relating to ability to bend the replaced knee. This is clearly an issue in development of knee prostheses.

Piloting of the questionnaire

In July 1989 I compiled a list of names and addresses of 21 of the patients under the care of surgeon S1. These patients had undergone knee replacement operation in the period between 1983 and 1989. Questionnaires were sent with a prepaid reply envelope and the covering letter shown in Appendix 6. It was explained to patients that they were able to comment on any aspects of the questionnaire which they found unclear. They were also assured that surgeon S1 and other doctors involved in caring for them would not see the replies to the questionnaire. This practice was continued for the rest of the study.

Of the 21 questionnaires, 13 were returned. A reminder and a further questionnaire elicited four further replies, thus gaining a response rate of 17/21 (81%). No comments were made by any respondents to the effect that the questionnaire was confusing so it was maintained in this form throughout the study.

Intended use of the questionnaire

The questionnaire was initially designed as an instrument for this thesis only. It was not intended as a portable outcome measure like the HAQ, AIMS, or WOMAC instruments. These were designed to assess the overall health impact of a general medical intervention, whereas the current instrument was intended to compare success between groups of patients in achieving the basic goals of knee replacement. If it were to be used for further research, the testing of a shorter version would be of interest. As it stands it is shorter than any of the published scales and elicits background information necessary for

comparative analysis.

CHAPTER 8. MAILING OF QUESTIONNAIRE AND RESPONSE RATES

Introduction

Chapter 6 described the way subjects were identified for study. This chapter describes the method for approaching these subjects, and reports the response rates obtained. Analysis was carried out to ascertain the reasons for and factors associated with non-response.

Mailing

The original plan was to send questionnaires to all patients listed except for those known to have died and those with obviously inadequate addresses. To those patients not replying within two weeks, a reminder letter was sent, and after two further weeks a further reminder and questionnaire was sent to non responders.

The timetabling of questionnaire mailing is shown in Table 8-1. It illustrates that after the first three series had been mailed, postcard reminders after a fortnight were dispensed with. This was because this practice did not appear to boost response rates very greatly and further complicated an already complex schedule.

The schedule involved the staggered mailing of questionnaires, starting with relatively small patient groups until the administration of the survey was running smoothly.

Response rates by series are shown in Table 8-2. Replies to the questionnaire were obtained for no more than 60% of all patients who had received a knee replacement. However 296 patients on the list are known to have died. Some of these were known about prior to mailing and others were discovered after the mailing. Similarly the addresses held on some patients were inadequate; in some patients this was so obvious that no questionnaire was even mailed while for others questionnaires were returned by the post office. For a further group the post office delivered questionnaires at a given address but later returned them because the named patient was "not known at this

address". The details are shown by series in Table 8-3.

Subtracting the numbers of patients shown in these five categories from the total patients left 1166 patients, from whom 960 replies were obtained (82%). This alternative response rate is shown by series in Table 8-4.

Analysis of factors influencing response rate

The difference in response rates between the series seen in Table 8-2 persisted even when dead patients and those of unknown address were excluded. Part of the reason for this may be the longer time periods over which some series were accumulated; for example Series 04 and 06 both began in 1973 whilst Series 01, 02, 03, 05, and 10 did not begin until the 1980s. Part of the poor overall response rate in Table 8-2 was because of deaths which happened more frequently for patients in series which began in the 1970s. Figure 8-1 shows how the series varied in the time since operation.

A very clear time effect emerged across all series in the response rate; the more recently that the operation was carried out, the higher the response rate. Table 8-5 shows the change in response rate according to year of operation. In Table 8-5 the denominator is taken to be all the non-excluded patients. This is also seen in Figures 8-2 and 8-3, where the denominator consists firstly of all patients in the original list, and secondly of all the non-excluded patients. The Figures only relate to the larger series which included patients operated upon before 1980.

The response rate was tabulated against age of patient, gender, and type of arthritis. Response rates were clearly higher among younger patients and lowest among the elderly. There was little difference according to gender, and a slightly higher response among OA patients compared with RA.

In order to investigate whether the difference in response rates among surgeon series was explained by differential distributions of age, gender, type of arthritis or year of operation, multifactorial analysis was carried out. This consisted of conventional logistic

regression (Cox 1970) carried out using the statistical package GLIM (Payne 1986). Models were fitted which assessed the effects of each factor when taken singly (unadjusted) and in combination with every other factor (adjusted).

The odds ratios for responding are shown below in Table 8-6. The analysis was carried out including all patients in the response rate denominator, and also excluding deaths and unknown addresses.

The difference in response rates between surgeon series persisted even when deaths and unknown addresses were removed from the denominator. A clear tendency is shown for response rates to improve according to how recently the latest operation was carried out. Response rates dropped with age, and RA patients were less likely to respond than OA patients. Odds ratios remained fairly similar in the two analyses for all five factors, although confidence intervals widened. The above analysis was based on "unifactorial" analysis, where each factor was assessed on its own. The analysis in Table 8-7 shows the **independent** effects of each factor once the possible confounding effects of the other factors were taken into account.

The effects of the various factors on response rate still seemed to persist (except for gender). It may be noted however that the difference between surgeon series, when allowing for the effects of the other four factors, and when excluding deaths and unknown addresses, was much less striking ($p=0.04$). Confidence intervals for odds ratios for the various series were wider in multifactorial analysis which excluded deaths and unknown addresses.

Discussion

Clearly older patients and patients operated upon in the 1970s were under represented in the questionnaire returns. Among patients operated upon in the 1980s there seems to be little difference in response rate among the surgeon series. Once deaths and unknown addresses were excluded, response rates exceeded 80%. It is of interest to note that differences according to age and year of operation persisted even when these exclusions

were made. This suggests that the 206 non responders in this survey for whom no reason for their non response was known may in fact have died or moved address (possibly into residential care). If resources allowed, the NHS Central Register could be asked to trace whether the patients had died, but this would depend on patients' having an NHS number. Alternatively non response may have occurred for patients with disabling arthritis in their hands, rendering them unable to write.

This analysis shows the difficulties of keeping track of a population undergoing surgery at an advanced age. Among those living patients whose addresses were up to date, a good response rate was obtained, suggesting that short term follow up studies by questionnaire are feasible. Longer term follow up, even when addresses are kept up to date, will be hampered by high mortality among this elderly population.

Table 8-1. Mailing timetable of survey

Series	1st mailing	Postcard reminder	2nd mailing	No. patients mailed
01-a	Jul 1989	Aug 1989	Aug 1989	21
01-a (1 year repeat)	Jul 1990			17
01-b	Apr 1990	Apr 1990	May 1990	46
01-b (1 month repeat)	May 1990			37
01-b (1 year repeat)	Apr 1991			
02	May 1990	May 1990	Jun 1990	65
03	Jun 1990	Jun 1990	Jul 1990	48
04	Sep 1990		Oct 1990	125
05	Aug 1990		Sep 1990	146
06	Oct 1990		Nov 1990	213

07	Nov 1990	Dec 1990	360
08	Jan 1991	Feb 1991	355
09	Jun 1991	Jul 1991	145
10	May 1991	Jun 1991	76
10 (1 month repeat)	Jun 1991	Jul 1991	46

Table 8-2. Response according to series and need for reminders

Series	No. on initial list	No. replying to:-			Total no. responders (% response)	No. who filled in "knee" items
		Initial q'aire	Reminder	Further q'aire		
01	67	35	8	11	54(81%)	54
02	65	49	5	3	57(88%)	52
03	48	28	2	2	32(67%)	31
04	125	50	-	14	64(51%)	60
05	146	83	-	15	98(67%)	94
06	213	93	-	10	103(48%)	93
07	360	211	-	25	236(65%)	225
08	355	157	-	39	196(55%)	183
09	145	73	-	1	74(50%)	73
10	76	40	-	6	46(58%)	42
Total	1600	819	15	126	960(60%)	907

* Knee questions are those including visual analogue lines.

Table 8-3. Reasons for exclusions by series

Series	No. on initial list	No. deaths known:-		No. incomplete addresses:-		Not known at address
		Before mailing	After mailing	Before mailing	After mailing	
01	67	0	0	2	0	1
02	65	0	4	0	0	1
03	48	10	3	0	1	0
04	125	10	18	0	1	4
05	146	17	7	2	1	2
06	213	24	18	14	3	16
07	360	35	11	22	10	12
08	355	72	24	2	3	18
09	145	20	4	9	4	8
10	76	13	6	0	0	2
Total	1600	201	95	51	23	64

Table 8-4. Response rate by series after making exclusions

Series	No. not excluded	Respondents	% response
01	64	54	84
02	60	57	95
03	34	32	94
04	92	64	70
05	117	98	84
06	138	103	75
07	270	236	87
08	236	196	83
09	100	74	74
10	55	46	84
Total	1166	960	82

Table 8-5. Response rate by series and year of operation : %(no.)

Surgeon Series	Year of most recent primary knee replacement				
	Not known	1973-9	1980-4	1985-7	1988-90
01	90(10)	-	75(8)	86(28)	83(18)
02	100(2)	-	100(1)	94(31)	96(26)
03	100(4)	-	100(3)	93(27)	-
04	80(5)	38(32)	93(30)	80(25)	-
05	83(47)	-	44(16)	91(23)	100(31)
06	74(19)	66(29)	70(44)	85(46)	-
07	79(19)	74(19)	84(95)	91(88)	96(49)
08	0(2)	64(28)	83(123)	92(77)	100(6)
09	100(1)	61(28)	74(47)	84(19)	100(5)
10	100(9)	-	81(32)	83(12)	50(2)
Total	82(118)	59(136)	80(399)	89(376)	95(137)

Table 8-6. Unifactorial analysis of response rate.

Factor	Category	Including all patients			Excluding deaths and unknown addresses		
		n	Odds ratio	95% confidence interval	n	Odds ratio	95% confidence interval
Year of operation	Not known	167	1.0	-	118	1.0	-
	1973-9	269	0.31	0.20-0.46	136	0.31	0.17-0.55
	1980-4	558	0.96	0.67-1.36	399	0.85	0.50-1.45
	1985-7	451	2.08	1.44-3.02	376	1.77	1.00-3.14
	1988-90	155	3.75	2.22-6.35	137	4.02	1.66-9.70
Surgeon series	01	67	1.0	-	64	1.0	-
	02	65	1.71	0.66-4.43	60	3.51	0.94-13.1
	03	48	0.48	0.21-1.13	34	2.96	0.62-14.1
	04	125	0.25	0.13-0.51	92	0.42	0.19-0.95
	05	146	0.49	0.24-0.99	117	0.96	0.41-2.20
	06	213	0.23	0.12-0.44	138	0.55	0.25-1.18
	07	360	0.46	0.24-0.87	270	1.29	0.60-2.76
	08	355	0.30	0.16-0.56	236	0.91	0.43-1.93
	09	145	0.25	0.13-0.50	100	0.53	0.23-1.18
	10	76	0.37	0.17-0.79	55	0.95	0.35-2.53
Sex	Male	394	1.0	-	278	1.0	-
	Female	1198	1.05	0.83-1.33	884	0.87	0.60-1.25

	Not known	8	0.69	0.17-2.80	4	-	-
Age	<55	91	1.0	-	87	1.0	-
	55-	335	0.90	0.49-1.65	305	1.27	0.63-2.56
	70-	676	0.45	0.26-0.79	570	0.66	0.35-1.25
	85-	111	0.25	0.13-0.48	86	0.37	0.17-0.79
	Not known	387	0.07	0.04-0.12	118	0.66	0.31-1.41
Type of arthritis	OA	701	1.0	-	512	1.0	-
	RA	466	0.73	0.57-0.92	315	0.73	0.51-1.05
	Other	65	0.74	0.44-1.23	45	0.68	0.32-1.42
	Not known	368	1.29	0.99-1.69	294	1.02	0.69-1.50

Table 8-7. Multifactorial analysis of response rate:

		Including all patients		Excluding all deaths and unknown addresses	
Factor	Category	Odds ratio	95% confidence interval	Odds ratio	95% confidence interval
Year of operation	Not known	1.0	-	1.0	-
	1973-9	0.09	0.05-0.18	0.16	0.07-0.37
	1980-4	0.22	0.12-0.40	0.38	0.18-0.82
	1985-7	0.41	0.23-0.71	0.71	0.34-1.50
	1988-90	0.75	0.38-1.48	1.82	0.66-4.97
Surgeon series	01	1.0	-	1.0	-
	02	0.43	0.15-1.29	3.42	0.84-13.9
	03	0.21	0.07-0.59	3.44	0.66-17.8
	04	0.11	0.04-0.29	0.50	0.19-1.35
	05	0.81	0.35-1.86	1.0	0.41-2.44
	06	0.09	0.04-0.23	0.52	0.20-1.37
	07	0.20	0.08-0.47	0.93	0.37-2.32
	08	0.14	0.06-0.35	0.80	0.31-2.08
	09	0.13	0.05-0.33	0.49	0.18-1.35
	10	0.12	0.05-0.32	0.83	0.27-2.50
Sex	Male	1.0	-	1.0	-
	Female	1.14	0.85-1.52	0.98	0.66-1.45

	Not known	0.75	0.13-4.32	-	-
Age	<55	1.0	-	1.0	-
	55-	1.08	0.57-2.06	1.25	0.59-2.65
	70-	0.47	0.25-0.86	0.55	0.27-1.11
	85-	0.35	0.17-0.73	0.39	0.16-0.93
	Not known	0.02	0.01-0.05	0.41	0.16-1.03
Type of arthritis	OA	1.0	-	1.0	-
	RA	0.70	0.51-0.96	0.76	0.50-1.17
	Other	0.84	0.43-1.63	0.59	0.26-1.33
	Not known	0.51	0.30-0.87	0.32	0.17-0.63

Figure 8-1. Distribution of most recent primary knee replacement by Surgeon series

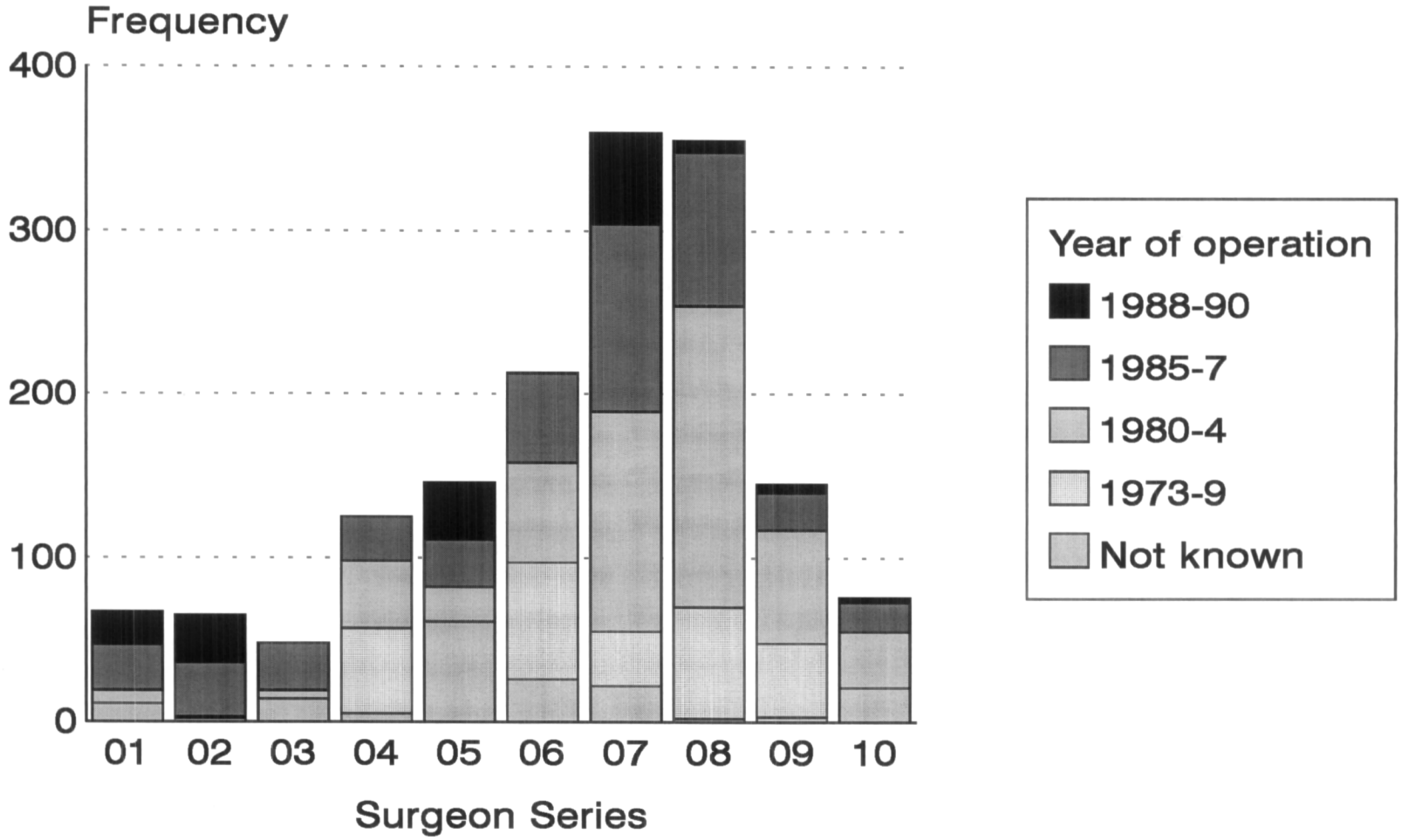


Figure 8-2. Response rate (a) versus year of most recent primary knee replacement.
For Series 04, 06, 07, 08, 09.

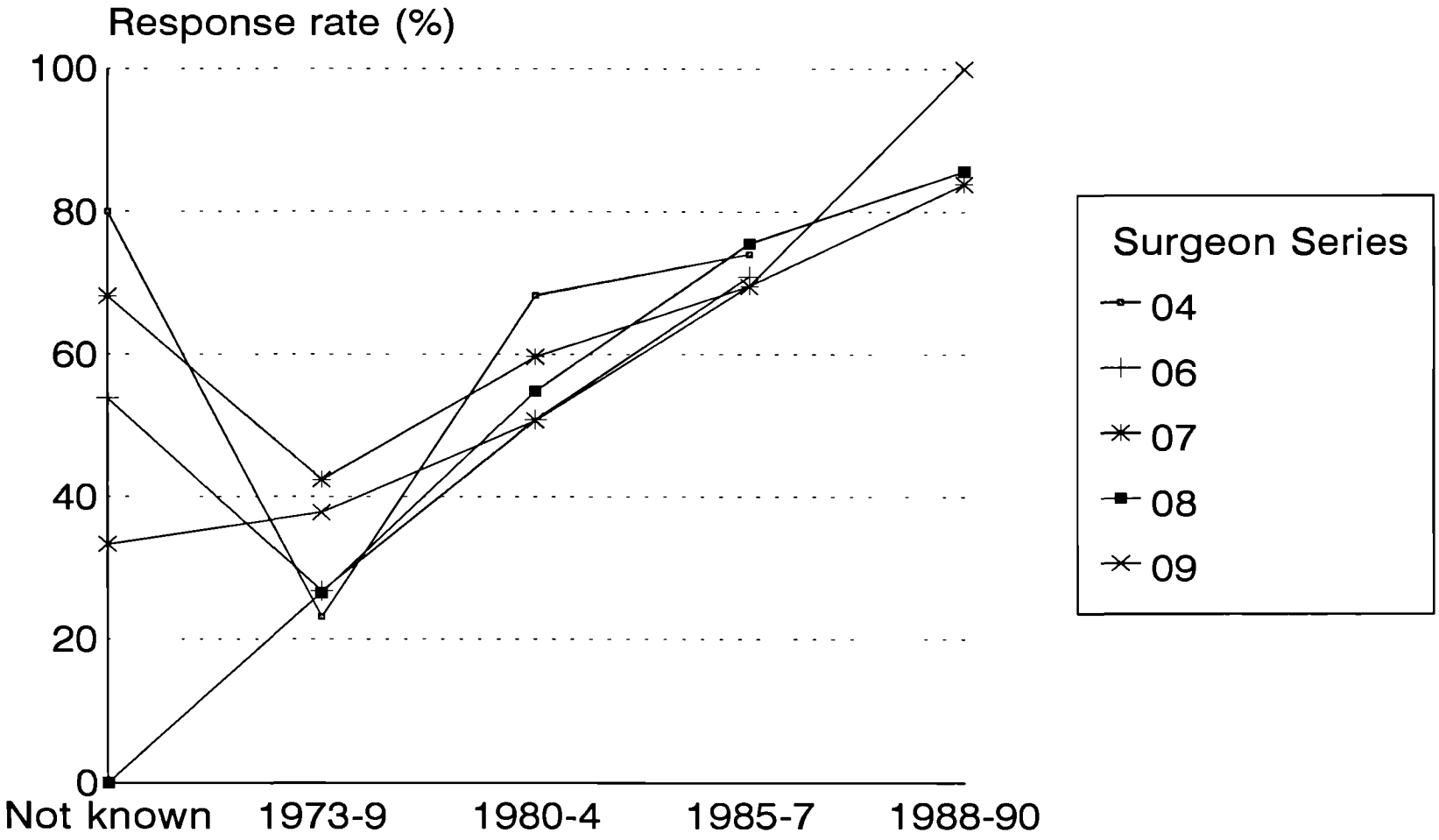
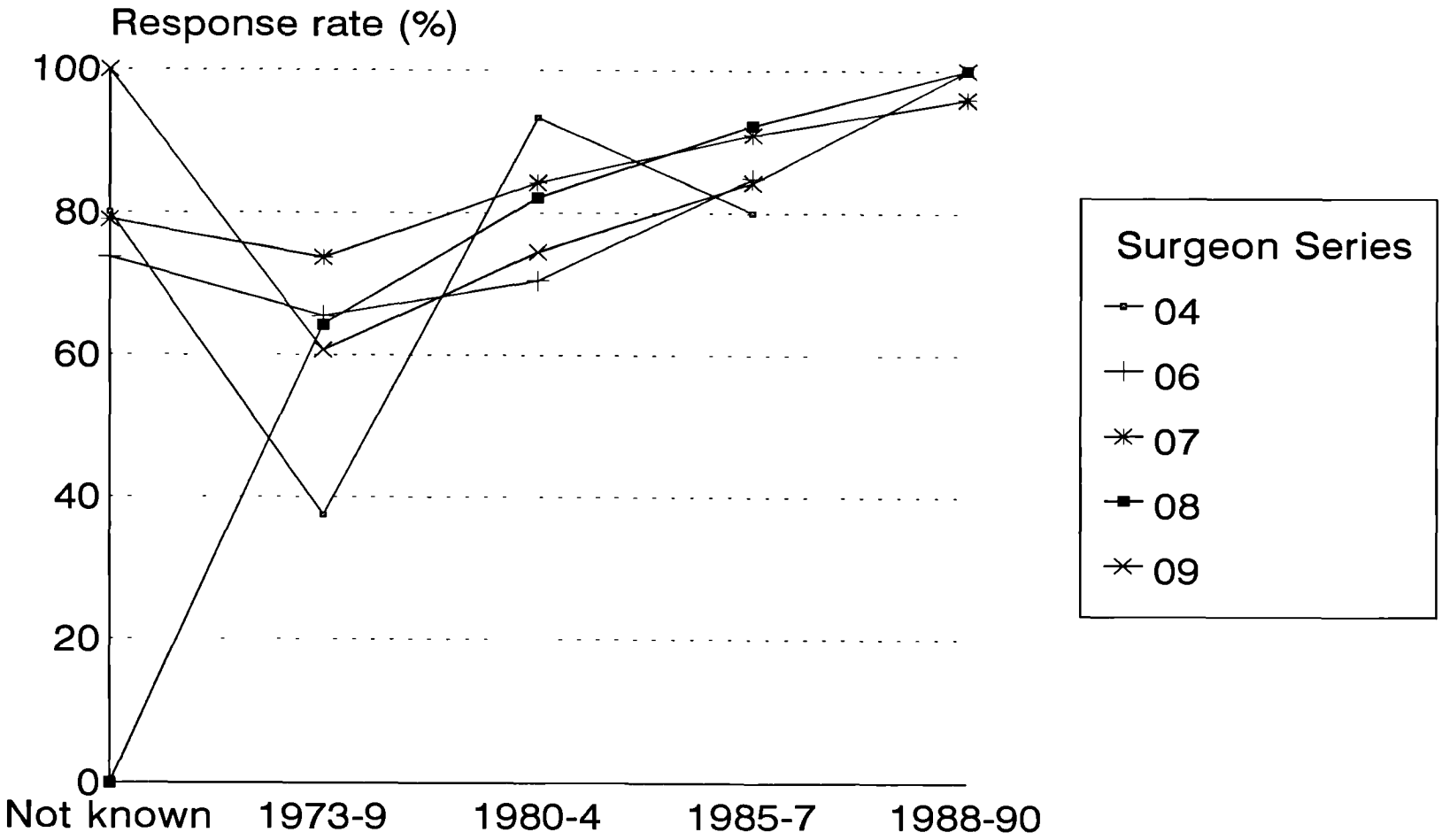


Figure 8-3. Response rate (b) versus year of most recent primary knee replacement.
For Series 04, 06, 07, 08, 09.



CHAPTER 9.
REPEATABILITY OF THE QUESTIONNAIRE

Introduction

Any measuring instrument is subject to several sources of variability other than true differences in the quantity measured. In the context of this questionnaire, there may be:

(i) Genuine variation in results obtained between patients

(ii) Genuine changes for any given patient from one occasion to another

(iii Inability of the patient to classify items reliably.

It may be very difficult to distinguish between sources (ii) and (iii), even conceptually (Fitzpatrick et al 1989). Sceptics may dismiss patients' expressed opinions as meaningless and impute any change of mind by a patient to source (iii) rather than source (ii). However one patient who responded to this survey mentioned that it was very common to have a good day and a bad day succeeding each other. This may be imputed to source (ii). A patient should ideally be enabled to take a broader view than simply being influenced by the problems experienced on the particular day when the questionnaire is filled in. If patients cannot classify their feelings reliably, this could be the fault of the questionnaire.

In connection with the broad aims of the research, quantifying "within patient" variability may be important for the following reasons:

When comparing clinicians' assessments with answers to postal questionnaires by patients, any differences may be simply a consequence of patient variability. However this variability should not happen in any systematic direction. Thus a global analysis which showed clinicians giving consistently more optimistic assessments than the patients

themselves is unlikely to be due to within patient variability.

If comparing responses between different prostheses to the questionnaire, once again within patient variability should not work in any systematic direction to make one prosthesis spuriously appear better. In this situation statistical analysis should focus on between patient variability, but if this is inflated by within patient variability, then genuine differences between prostheses may be more difficult to demonstrate (Bellamy and Campbell 1989).

Thus a special sub-study was carried out to estimate "within patient" variability in responses to the questionnaire. The study can only touch upon the reliability of the instrument as opposed to its validity. Concerning validity, it is not possible to test "criterion validity" since no gold standard exists. "Construct validity" will be examined in later chapters, especially Chapter 11, where the relationship between this questionnaire and clinicians' assessments will be investigated.

Materials and Methods

These included all patients in Series 01 (n=67) and those in Series 10 (n=76). From Series 01, 22 were first sent a questionnaire in July/August 1989 (Series 1a), and 45 in April 1990 (Series 1b). Of those who replied to the questionnaire in Series 1b, a second questionnaire was mailed one month after receipt of the first. For both Series 1a and 1b, a further questionnaire was sent one year after the first mailing. For Series 10, a second questionnaire was sent one month after obtaining a reply to the initial questionnaire.

Separate comparisons were made between replies for patients who returned two questionnaires one month apart, and between replies for patients who returned questionnaires one year apart. It was considered that one year was a sufficient time for detecting changes which may occur for patients in their underlying condition. In one month however no extensive changes should occur in the status of the knee replacement and any variability should ideally be small relative both to variability within patients over

the course of one year and to variability between patients.

Statistical methods

1. Categorical variables

Various measures of agreement between categorical variables have been proposed. When comparing responses to a two category variable obtained on two occasions, any systematic change may be assessed using McNemar's chi-squared test, while the relative degree of agreement is given by the Kappa statistic (Fleiss 1981). If the variable is ordinal and has more than two categories then a "quadratically weighted" kappa statistic may be used, so that misclassifications of varying seriousness may be given due weighting. This modified kappa statistic bears a strong analogy to the intraclass correlation coefficient which has been used for quantitative variables. Quadratically weighted kappa has been used in this chapter rather than ordinary kappa for all ordinal items.

A further statistic (π) has been suggested by Chinn and Burney (1987) namely the "average correct classification rate", which measures the absolute agreement. This was regarded as a useful complement to the kappa statistic which measures relative agreement.

Knee oriented questions

For the visual analogue items we have scores from 0 to 6 for each question. The number of categories was large in relation to the number of subjects and calculating kappa statistics was not desirable. Instead the scores were regarded as arising from a continuum and agreement was assessed using the within subject standard deviation and the intra class correlation coefficient (Snedecor and Cochran 1980). This may also be calculated for the total knee score defined as:

Item1 + Item2 + Item3 + Item4 + Item5 + (6 - Item6a) + (6 - Item6b)

Such a score, used later in the main analysis, ranges from a minimum of 0 to a maximum of 42.

To investigate further the repeatability of the total score, the graphical technique of Bland and Altman (1986) was used.

The observational unit for this analysis was the patient rather than the knee, as in Chapter 10. For each knee item and the total knee score, the average for left and right knees was used for patients who had both knees replaced. For the seventh and eighth items of the knee oriented parts of the questionnaire, results were analysed for left and right knees separately but since these differed little, results for left knees only are presented.

RESULTS

Response rate

In Series 1a, 22 questionnaires were mailed on the first occasion and 17 replies were received. One year later, of the 17 mailed, 16 replies were received.

In Series 1b, 45 questionnaires were mailed and 37 were received back. One month later, these 37 received a questionnaire of whom 26 replied. One year after the first mailing, 26 received a third questionnaire and 22 replied.

In Series 10, 76 questionnaires were mailed and 45 were received back. One month later these 45 received a further questionnaire, and 25 of these replied.

Hence the overall response rate to a first mailing was 99 out of 143 (69%), to a mailing one month later 51 out of 82 (62%), and to a mailing one year later 38 out of 43 (88%). Thus out of 121 who were potentially eligible to reply initially and one month later, 51 actually did so (42%); and out of 67 eligible to reply initially and one year later, 38 did so (57%).

Thus to evaluate "short term repeatability", the available number of subjects was 51 (68 knees) and for "long term repeatability" it was 38 (46 knees).

Unfortunately not all subjects answered every question, so sample sizes for analyses of some questions are less than the number quoted.

Patient oriented questions

Short term repeatability

Cross tabulations for each item of the general parts of the questionnaire were made and summary statistics for each item are shown in Table 9-1.

Broadly speaking each measure of agreement on the categorical variables resulted in similar rankings of the 30 items shown above. According to the arbitrary labels given to values of Kappa by Landis and Koch (1977), almost perfect agreement was obtained for the sociodemographic variables and information on most of the joints replaced. Since these were factual items, such good agreement was to be expected. Substantial agreement occurs for information on aids used; perhaps the use of three categories made these items less than completely objective. The items pertaining to functional activities resulted in moderate to substantial agreement. The items which resulted in least agreement were rising from a chair and getting out of bed. Only moderate agreement was obtained with items concerning pain in various joints.

Out of the sample size of 51 available, complete data for the two occasions were available for most items for at least 45 cases. Items on joint pain were however badly answered, as were questions on which people lived with the respondent. It is possible that respondents did not tick items which they considered not applicable (ie the true answer should have been "no").

It may appear odd that any disagreements arose at all concerning the factual items. This seemed to arise in the multiple items relating to one theme (eg which people live with

you) where answers had to be ticked "No" or "Yes" separately. Perhaps a tick in the box labelled "No" was meant to indicate "Yes" while an unticked item was meant to indicate "No". It is also possible that people living with the respondent may have changed over the course of the month.

Concerning the knee items represented as visual analogue items, the first three items showed better agreement than the last four. This could be attributable to fatigue on the part of the patient. However the last two items concerning pain required alertness on the respondent's part since a good outcome involved marking the left hand end of the scale , whereas the first five items involved the opposite. There was evidence that three of the patients had been caught out by this; answers to their first five items had suggested a good outcome (right hand end of scale ticked) but they had also ticked the right hand end of the scale for pain items (supposedly indicating very bad pain).

Analysis of variance was carried out for knee items 1 to 6b (visual analogue lines), and gave within subject standard deviations and intraclass correlation coefficients (as seen in Table 9-2).

The total score, with a maximum range of 0 to 42, had a within subject standard deviation of 3.61. Figure 9-1 shows the differences which seemed least when the patients returned high scores. The intraclass correlation suggests that sample sizes for future studies which use this knee score as chief outcome measure should be inflated by 12% over what would be necessary if no responder error existed.

Long term repeatability

The cross tabulations produced for each item of the general parts of the questionnaire were summarised as before and are shown in Table 9-3.

As for the results obtained for the analysis of short term agreement, each measure of agreement results in similar rankings of the 30 items shown. Once again, the best agreement was obtained for the sociodemographic variables and information on which

joints had been replaced (the factual statements). Substantial agreement occurred once again for information on aids used. Items concerning pain experienced in various lower limb joints resulted in moderate agreement and similarly for those items relating to functional activity. The items which resulted in least agreement were rising from a chair (as for short term repeatability), right knee and right ankle/foot pain.

Once again occasional disagreements occurred with the factual items on the questionnaire. This again arose in the multiple items relating to one theme where answers had to be ticked "No" or "Yes" separately. Over the period of one year it is possible that such things as marital status could genuinely have changed (one subject responded as married on the first occasion and as widowed one year later). For this analysis all items may well involve true within subject variation as well as errors in interpreting and completing the questionnaire itself.

Since there may have been systematic within subject changes over the course of one year, McNemar's test was carried out to investigate whether there was a tendency for the change to occur in one particular direction. No evidence of this was found for any of the items, probably because observed changes were in any case rare, and the sample size was fairly small.

Out of the sample size of 38 available, complete data for the two occasions was available for most items for at least 34 cases. The exceptions related to questions on pain experienced in various joints, and items concerning who the respondent lived with. If anything, data completeness from these respondents was greater than for that obtained from the respondents in the short term repeatability analysis.

Analysis of variance carried out on knee items 1 to 6b (visual analogue lines) gave within subject standard deviations and intra class correlation coefficients as shown in Table 9-4. Consideration of the knee items suggests that several vary little over the course of one year. The total score has a very high intra class correlation, suggesting that the instrument may be relatively insensitive to changes. Figure 9-2 shows the differences, which were again small for high scores. The mean difference was 1.2 (95% confidence

interval -0.2 to 2.6, p=0.09) giving weak evidence of a small decline in satisfaction over one year.

Several of the items , and the total score, seemed to have **lower** variability over one year than over one month. This is likely to be a random finding but suggests that true changes for individual patients over the period of a year are unlikely to be easily detected with this index.

Comparison of one month and one year repeatability

In the tables shown above it appeared that the Kappa values for one month repeatability exceeded that for one year repeatability for 20 of the 30 items. For nine items the reverse was the case, whilst one item (whether the respondent lived with a friend) gave equal Kappa values for the two situations.

For most of the functional questions there was closer agreement over one month than over one year. The difference was particularly striking for the items concerning rising from a chair, and for getting up or down stairs, and cutting toenails. Conversely there was little difference concerning dressing, getting out of bed or picking items off the floor. It would appear that changes over one year were greater for activities where the subject has some choice than for activities which have to be carried out anyway.

Questions on the use of aids, sociodemographic variables and whether various joints had been replaced did not seem to differ markedly in their repeatability over one month compared to over one year.

Concerning the level of pain reported in the various joints, this seemed to vary less over one month than over one year for the knee joints and the ankle joints. Variation for hip joint pain did not follow such a pattern, however.

As was remarked above, the knee items appeared to demonstrate **less** variability over the

course of one year than over one month. This scarcely seems logical and is presumably a feature of the partially differing samples of subjects used for the two analyses. In either case the total score appears to demonstrate good repeatability and thus to be of potential use in detecting differences between subgroups of patients.

Discussion

The following points seem to emerge from this study of the reliability of the questionnaire.

1. Patients' responses to factual questions seem to be reliable, as one would hope. Patients who managed to reply to this questionnaire twice are unlikely to be affected by mental confusion. The reliability of this questionnaire may therefore be overestimated by this sample of patients.

2. The repeatability indices for items concerning the use of aids was moderately good. Patients may have been unable to discriminate between using aids "sometimes" as opposed to "often". In studies which attempt to elicit the use of aids as an index of disability, these questions may need rewording.

3. Repeatability on items concerning joint pain were moderate- this may well reflect the intermittent nature of joint pain. Many sophisticated instruments to measure pain have been devised (Jadad and McQuay 1993) but involve far too many items for the purposes of comparing outcome of knee replacement in large groups of patients.

4. Repeatability for items pertaining to function was moderate. Some items varied more over 1 year than over 1 month and these may have potential for identifying real change within patients over a long period.

5. While some of the visual analogue items in the knee questionnaire demonstrated only moderate repeatability, the total knee score had a high intra class correlation

coefficient and appears to be of potential use in comparing outcome between groups of patients. Some patients clearly marked the scale at the wrong end for some of the items and future questionnaires may benefit from a little extra explanation.

Table 9.1 Agreement statistics for items of questionnaire derived from responses made on two occasions one month apart

Item (no. of categories)	n	Estimate (standard error) Proportion in agreement	Average correct classification	Weighted Kappa value
Functional activities:				
1. Dressing(3)	50	0.92(0.04)	0.96(0.02)	0.82(0.11)
2. Rising from chair(3)	51	0.78(0.06)	0.88(0.04)	0.63(0.10)
3. Going up stairs(5)	49	0.88(0.05)	0.94(0.03)	0.92(0.04)
4. Going down stairs(6)	50	0.80(0.06)	0.89(0.03)	0.84(0.06)
5. Walking distance(6)	49	0.74(0.06)	0.86(0.04)	0.91(0.03)
6. Getting out of bed(3)	51	0.73(0.06)	0.85(0.04)	0.48(0.14)
7. Picking up objects(4)	50	0.74(0.06)	0.86(0.04)	0.66(0.12)
8. Household chores(4)	50	0.72(0.06)	0.84(0.04)	0.85(0.04)
9. Cutting toe nails(4)	44	0.84(0.06)	0.92(0.03)	0.94(0.02)
10. Use of aids:				
a. Cane(3)	45	0.84(0.05)	0.92(0.03)	0.82(0.08)
b. Crutches(3)	45	0.96(0.03)	0.98(0.02)	0.73(0.19)
c. Frame(3)	42	0.98(0.02)	0.99(0.01)	0.85(0.06)
11. Joints replaced:				
a. Right knee(2)	48	0.92(0.04)	0.96(0.02)	0.82(0.09)
b. Left knee(2)	49	0.90(0.04)	0.95(0.02)	0.75(0.11)
c. Right hip(2)	46	0.96(0.03)	0.98(0.02)	0.87(0.09)

d. Left hip(2)	47	1.0(-)	1.0(-)	1.0(-)
e. Other joint(2)	46	0.96(0.03)	0.97(0.02)	0.54(0.24)
12. Joint pain:				
a. Right knee(3)	37	0.70(0.08)	0.83(0.05)	0.59(0.13)
b. Left knee(3)	30	0.87(0.06)	0.93(0.04)	0.88(0.05)
c. Right hip(3)	23	0.78(0.09)	0.88(0.05)	0.57(0.20)
d. Left hip(3)	20	0.75(0.10)	0.86(0.06)	0.50(0.20)
e. Right ankle(3)	30	0.77(0.08)	0.87(0.05)	0.79(0.07)
f. Left ankle (3)	27	0.85(0.07)	0.92(0.04)	0.77(0.13)
Sociodemographic variables:				
a. Marital (5)	47	1.0(-)	1.0(-)	1.0(-)
b. Housing(5)	48	0.96(0.03)	0.98(0.02)	0.92(0.06) *
c. Live with:				
Husband /wife(2)	46	0.94(0.04)	0.97(0.02)	0.87(0.07)
Bro/Sister (2)	35	1.0(-)	1.0(-)	1.0(-)
Son/ daughter(2)	42	1.0(-)	1.0(-)	1.0(-)
Other relative (2)	35	1.0(-)	1.0(-)	1.0(-)
Friend (2)	35	1.0(-)	1.0(-)	1.0(-)
Knee items (non visual analogue)				
7. Can bend (2)	45	0.87(0.05)	0.93(0.03)	0.70(0.11)
8. How long can you bend? (4)	46	0.72(0.07)	0.84(0.04)	0.78(0.08)

(-) denotes that the standard error could not be calculated because the index was equal

to one

* is for unweighted kappa since the variable was not ordinal

Table 9.2 Agreement statistics for visual analogue scale items derived from responses made on two occasions one month apart

Item:	Within subject standard deviation	Intraclass correlation
1. Compared with before operation	0.75	0.85
2. Improvement as expected	0.68	0.89
3. Compared with 1 year ago	0.85	0.78
4. Difference to overall health	1.27	0.64
5. Worthwhile	0.91	0.75
6a. Resting pain	1.17	0.62
6b. Moving pain	1.05	0.72
Total score	3.61	0.89

Table 9.3 Agreement statistics for items of questionnaire derived from responses made on two occasions one year apart

Item (no. of categories)	n	Estimate (standard error) Proportion in agreement	Average correct classification	Weighted Kappa value
Functional activities:				
1. Dressing(3)	38	0.90(0.05)	0.95(0.03)	0.89(0.05)
2. Rising from chair(3)	38	0.66(0.08)	0.80(0.06)	0.39(0.14)
3. Going up stairs(5)	36	0.75(0.07)	0.86(0.04)	0.71(0.12)
4. Going down stairs(6)	37	0.65(0.08)	0.80(0.05)	0.45(0.22)
5. Walking distance(6)	38	0.50(0.08)	0.69(0.06)	0.81(0.05)
6. Getting out of bed(3)	38	0.71(0.07)	0.84(0.05)	0.55(0.13)
7. Picking up objects(4)	37	0.70(0.08)	0.83(0.05)	0.56(0.15)
8. Household chores(4)	38	0.68(0.08)	0.82(0.05)	0.67(0.14)
9. Cutting toe nails(4)	35	0.63(0.08)	0.78(0.07)	0.54(0.15)
10. Use of aids:				
a. Cane(3)	36	0.75(0.07)	0.89(0.04)	0.85(0.06)
b. Crutches(3)	32	0.94(0.04)	0.97(0.02)	0.78(0.17)
c. Frame(3)	32	0.94(0.04)	0.98(0.02)	?
11. Joints replaced:				
a. Right knee(2)	38	0.90(0.05)	0.94(0.03)	0.79(0.10)
b. Left knee(2)	37	0.92(0.04)	0.96(0.03)	0.80(0.11)
c. Right hip(2)	34	1.0(-)	1.0(-)	1.0(-)

d. Left hip(2)	35	0.97(0.04)	0.99(0.02)	0.89(0.11)
e. Other joint(2)	35	0.97(0.03)	0.99(0.02)	0.79(0.21)
12. Joint pain:				
a. Right knee(3)	27	0.59(0.09)	0.75(0.08)	0.39(0.19)
b. Left knee(3)	27	0.70(0.09)	0.83(0.06)	0.65(0.11)
c. Right hip(3)	20	0.80(0.09)	0.89(0.05)	0.77(0.10)
d. Left hip(3)	19	0.74(0.10)	0.85(0.07)	0.46(0.28)
e. Right ankle(3)	22	0.64(0.10)	0.78(0.08)	0.36(0.18)
f. Left ankle (3)	21	0.57(0.11)	0.73(0.09)	0.42(0.19)
Sociodemographic variables:				
a. Marital (5)	34	0.94(0.04)	0.97(0.02)	0.91(0.06) *
b. Housing(5)	34	0.94(0.04)	0.97(0.02)	0.88(0.08) *
c. Live with:				
Husband /wife(2)	36	0.94(0.04)	0.97(0.02)	0.89(0.08)
Bro/Sister (2)	23	0.96(0.04)	0.98(0.02)	0.65(0.33)
Son/ daughter(2)	27	0.96(0.04)	0.98(0.02)	0.91(0.08)
Other relative (2)	22	0.96(0.04)	0.98(0.02)	0.65(0.33)
Friend (2)	21	1.0(-)	1.0(-)	1.0(-)
Knee items (non visual analogue)				
7. Can bend (2)	35	0.86(0.06)	0.92(0.04)	0.69(0.13)
8. How long can you bend? (4)	33	0.64(0.08)	0.79(0.06)	0.69(0.11)

(-) denotes that the standard error could not be calculated because the index was equal to one

* means that unweighted kappa was calculated since these items are not ordinal

? means kappa could not be calculated since only one of the two columns contained observations

Table 9.4 Agreement statistics for visual analogue scale items derived from responses made on two occasions one year apart

Item:	Within subject standard deviation	Intraclass correlation
1. Compared with before operation	0.40	0.95
2. Improvement as expected	0.73	0.86
3. Compared with 1 year ago	0.89	0.74
4. Difference to overall health	1.36	0.68
5. Worthwhile	1.10	0.57
6a. Resting pain	1.10	0.65
6b. Moving pain	0.89	0.79
Total score	2.96	0.93

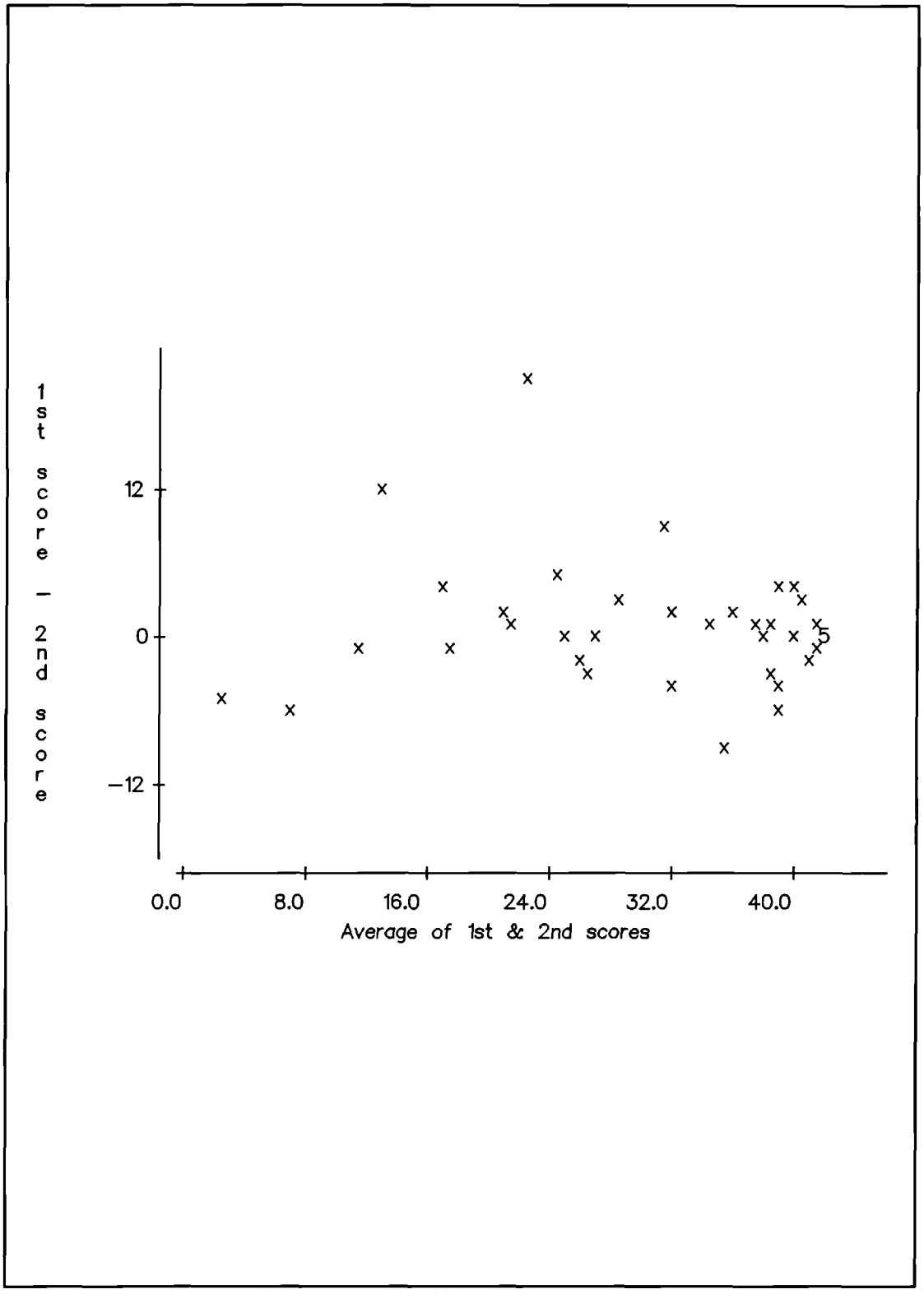


Figure 9-1. "Bland & Altman" plot of difference in scores obtained on two occasions one month apart

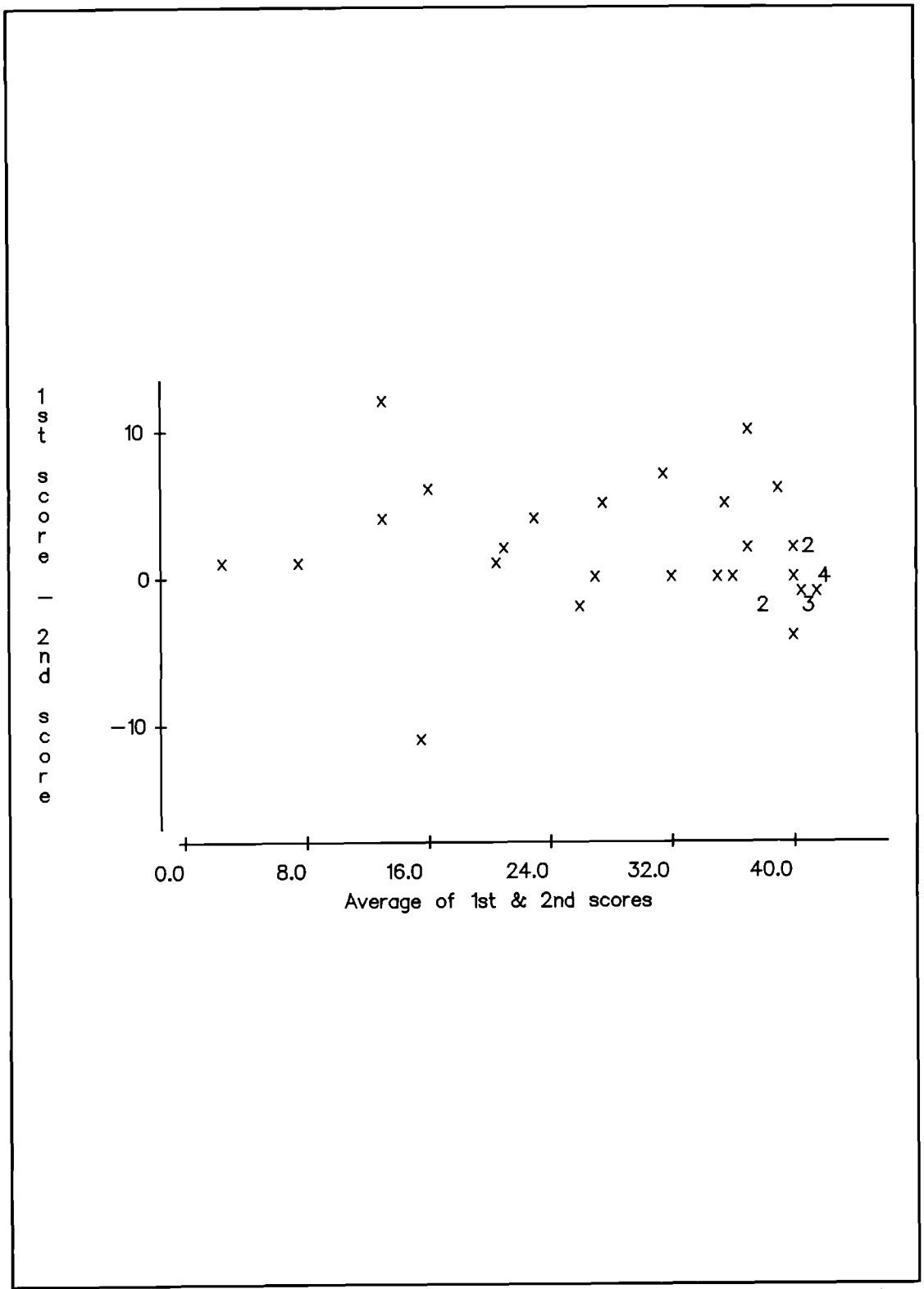


Figure 9-2. "Bland & Altman" plot of difference in scores obtained on two occasions one year apart

CHAPTER 10. RESULTS OF QUESTIONNAIRE SURVEY ACCORDING TO SURGEON SERIES

INTRODUCTION

This chapter describes general results of the survey in terms of the distribution of responses to various items of the questionnaire, and a detailed analysis primarily of the "knee score" is shown. This analysis was to ascertain whether differences in the distribution of patient satisfaction as seen in this knee score existed between surgeon series. Further analysis was also carried out to discover how the estimation of differences was affected by confounding variables. Secondary outcome variables such as a function score and existence of positive or negative comments by patients were subjected to similar analysis. The results are considered in the context of the differences in successful outcome according to prosthesis used as well as the surgeon who used the prosthesis.

METHODS

The chief outcome measure

The visual analogue items which consisted of direct questions about the perceived success of the knee replacement were more natural outcome measures than the functional items, since the latter could be related to a variety of problems in other joints or other medical diseases. The seven visual analogue items were combined into an overall score simply by adding. (After transforming the pain items from a scale of 0=good to 6=bad by setting pain=6-pain, these items accorded with items 1-5 in having 0=bad through to 6=good). By adding these scores, a total knee score was produced for which a maximum of 42 suggested the patient was perfectly satisfied with the knee replacement, while a minimum of 0 implied total dissatisfaction.

Objections may be made to the simple adding of different items. Bellamy and Campbell (1989) state themselves opposed to the aggregation of different dimensions of rating scales. However Bebbington (1977) argued that any reasonable set of weights can be

chosen arbitrarily without affecting the properties of the resulting index. Bebbington justified this case on a measure of personal dependency among impaired or handicapped individuals in a London borough, and a domestic self care measure among elderly people in Glamorgan. To investigate the validity of this approach for the current study, a principal components analysis of the seven items on the questionnaire was carried out to indicate an optimal weighting of the seven items when combining them into an overall score. This was carried out on complete information on 1178 knees. The first principal component was

0.42 x item1 + 0.41 x item2 + 0.39 x item3 + 0.36 x item4 + 0.40 x item5 - 0.32 x item6a - 0.35 x item6b.

This component explained 69% of the total variance between the 1178 responses, and the similarity of the coefficients for the various items suggests that a score derived by simple adding was unlikely to misrepresent the data to any important degree. (Negative coefficients for items 6a and 6b were because the pain items were untransformed from their original status of 0=bad to 6=good).

It was considered whether item3 ("How do you feel about this knee compared with one year ago?") should be omitted from the total score since it was possibly ambiguous to patients. However since the correlation of total scores which included and excluded item3 exceeded 0.99,it was retained in the total score.

This total score will from hereon be known as the "**knee score**".

Function score

A function score was derived as the second outcome variable. The restoration of function is one of the goals of knee replacement although it would be influenced by more factors than the success of the knee replacement alone.

The nine items at the start of the questionnaire all consisted of ordered categorical

variables. Each category of a particular item was assigned an integer quantity (for example, item 1: "How do you dress yourself?" gave 1=Not without someone's help, 2=By myself with difficulty, 3=By myself without difficulty). Then the items were combined as follows:

(item1 - 1)/2 + (item2 - 1)/2 + (item3 - 1)/4 + (item4 - 1)/5 + (item5 - 1)/5 + (item6 - 1)/2 + (4 - item7)/3 + (4 - item8)/3 + (4 - item9)/3

Each item was given equal weight in this combination by dividing the relevant component by the number of categories in the item. This was to prevent items with most categories having undue weight. Allowance was also made for the fact that some items were scored such that higher values implied better function while other items were scored the opposite way round.

This function score then had a theoretical range of 0 (very bad) to 9 (very good).

A principal components analysis was carried out and this suggested a simple adding of the first six items and subtracting of the last three (after scaling) would explain 62% of the variance.

Comment data

Patients were invited to make comments about their knee replacement. All questionnaires available in Series 01 and 05 (over 150 questionnaires) were scanned for comments which were listed. Themes were thus defined. Each patient's comments were coded for presence or absence of each theme.

1. Positive comments. These included expressions of gratitude for the operation, to the surgeon and mention of successful outcomes such as alleviation of pain or rediscovered ability to do mundane jobs or enjoyable activities.

2. Negative comments about pain still experienced. This was counted even if the

patient was also positive about the improvement achieved.

3. Negative comments about degree of knee flexion that the patient was (un)able to achieve. These included implicit statements such as being unable to pick objects up.

4. Other negative comments which did not pertain to pain or flexion. These may have included patients' fears that the joint would wear out.

5. Comments made by the patient about aspects of their medical history which might explain why they were not deriving maximum benefit from the knee replacement. This often included arthritis in the other knee or hips.

6. Comments that were irrelevant, such as the patient's general life history! These were only coded if nothing of what was written fell into categories (1) to (5).

The existence of a second, third or fourth type of comment was combined to denote presence or absence of any negative comment. This, together with the presence or absence of positive comments, comprised the data used for formal analysis.

Surgeon series to be compared

Descriptive data will be shown for all ten series. Also the assessment of the "surgeon series" effect will primarily rely on all ten series. However when overall series differences were identified, formal analysis was restricted, in order of importance, as follows

1. Comparison of Series 04+06, 07, 08, 09. These involved the four chief surgeons with the longest history of carrying out knee replacements, who comprised the initial study plan. Each surgeon had a different prosthesis (see Chapter 6).

2(a) Comparison of Series 08 with 05. Series 05 involved the same prosthesis as series

08, but the surgeon in series 08 was the prosthesis inventor.

2(b) Comparison of part of Series 07 with series 10. These involved the same prosthesis but the surgeon for series 07 was the inventor. Only part of Series 07 was included because the surgeon radically modified the prosthesis, and only patients who underwent knee replacement with the older style were included. All the patients in series 10 had the older form of the prosthesis.

Both comparisons 2(a) and 2(b) were to explore the possibility that the prosthesis performed best in the hands of the inventor.

3. Comparison of Series 01 and 02. These involved the same prosthesis (invented and first used in USA) but neither surgeon was the inventor. If differences were found, it would suggest a surgeon effect would be suggested even if one surgeon was not the inventor.

4. Comparison of series 04 and 06. These involved the same prosthesis and the same surgeon, but at two different hospitals. Any differences found could have been due to patient types (selection criteria for surgery are unlikely to have differed), social environment or rehabilitation policy.

Multifactorial analysis

The following variables were included on 907 patients: surgeon series, gender, diagnosis, number of knees operated upon, knee score, age at operation, time since operation, marital status, housing status, use of aids, negative comments, positive comments and functional score. These were calculated as follows:

Surgeon series: ranged from 01 to 10 as described in Chapter 6.

Gender: male, female, unknown.

Diagnosis: OA, RA, Other, Unknown

No. knees operated upon: Left, Right, Both.

Knee score: Computed as described, but for patients with both knees replaced, the average was taken. Complete data were only available for 819 patients.

Age at operation: Again, for patients with bilateral replacements, the average age at the two operations was computed. This was then categorised for all patients as under 55, 55-64, 65-74, over 75, unknown.

Time since operation: This was taken as the time gap between operation and sending of the questionnaire. For bilateral replacements, the average of the two was taken. The result was categorised as under 3 years, 3-6 years, 6-10 years, over 10 years, unknown.

Marital status: Married, single, widowed, divorced/separated, other, unknown.

Housing status: Council rented, privately rented, housing association, owner occupier, old people's home, unknown.

Use of aids: None used; cane, crutches or frame sometimes used (not often); cane, crutches or frame often used.

Negative comments: None, some.

Positive comments: None, some.

Functional score: Computed as above and remaining ungrouped. Complete data were available for 807 patients.

Unifactorial comparison

Statistical methods based on the Normal distribution were unlikely to be justified. For comparison of knee score between several groups of patients, the non parametric Kruskal Wallis test was used.

Multifactor comparison

It was likely that different sorts of patient were selected for surgery in the various series since they were not jointly participating in any organised multicentre study. Thus data on factors listed above needed to be included in a multifactorial analysis. Other orthopaedic variables on the patient's preoperative state (eg degree of deformity in the angle of tibial shaft to femoral shaft) would have been desirable, but not all surgeons measured this and even those who did seemed to differ in their method of measurement. Hence only the factors listed above may be considered unambiguous though it is accepted that these alone cannot capture all facets of the different patient populations.

Regression analysis

Multifactorial analysis of this nature could not be carried out with non parametric methods. In addition, some method was required where the magnitude of differences according to surgeon series could be estimated, with or without adjustment for confounding factors. Thus two approaches were considered.

(i) Use of an ordinary multiple regression model on the outcome variable. The assumptions are made that "residuals" (ie the difference between a patient's observed score and that predicted from the multiple regression model) follow a Normal distribution and their variance is constant. This approach was reasonable for the function score but not the knee score (which followed a negatively skewed distribution).

(ii) Regarding the outcome variable as binary. Then it would be possible to model the probability of a good result according to the surgeon series, the age, sex and diagnosis of the patient, and time since operation (the so-called explanatory

variables). The logistic regression model was used (Cox 1970), which allows comparisons between subgroups of patients by calculating the odds ratio of a poor result.

This approach was also used for the knee score by dichotomising it into a "good/poor" subdivide, by calling all values above a given threshold "good" and all those below "poor". The disadvantage here was that data was wasted in grouping patients into a simple good / poor classification, and the statistical power was thus reduced. The question also arose as to where to set the threshold; in fact it was set at the median score of 35 to provide the greatest statistical power possible under such a scheme. Thus all values of 35 and above were classed as "good" and those below 35 as "poor".

In either of the two regression models shown above, analysis could be carried out using the statistical package GLIM (Payne 1986). Different numbers of factors could be included. The simplest just included a single explanatory factor (eg surgeon series), to provide odds ratios of a good result for one subgroup when compared with another. A more complex model involved all other factors also so that adjusted odds ratios could be calculated for each subgroup of patients. Ninety five per cent confidence intervals for all odds ratios are presented.

Patients with both knees replaced- knee score

A separate analysis for left knees and right knees was not intuitively sensible, yet the usual approach in the orthopaedic literature (see chapter 3) of regarding every operated knee as providing an independent observation is not statistically sound. An intermediate approach involved taking the average of the total scores for the right and left knees. This was also done for age at operation and time between operation and sending the questionnaire. Hence even patients with both knees replaced only provided a single data point for analysis. This may seem a little unfair in therefore assigning too little weight to patients who had both knees replaced, so the technique of weighted regression was employed whereby observations on "bilaterally replaced" patients were given extra

weight, in a manner described in Appendix 7. The "degrees of freedom" needed for statistical tests was based on the number of patients, not the number of knees.

All multifactorial analyses require data to be present for every factor included in the model. Unfortunately data were missing for a number of the factors, and so the data set available for multifactorial analysis was reduced.

RESULTS

Overall results

The questionnaire and the distribution of responses to each item are shown in Appendix 8, but an overview of these results follows.

Functional items

In every item each option was chosen sufficiently frequently that no option seemed redundant. In most items there was one option which was clearly ticked most frequently, patients being most likely to say :

(1) They dressed themselves without difficulty

(2) They could stand up unaided if they held on

(3) They climbed and (4) descended the stairs holding the bannister and leading with the same leg

(5) They could walk for up to 10 minutes, or from 10 to 30 minutes (these two options were chosen with roughly the same frequency).

(6) They got out of bed without help with or without holding on (these options were of equal frequency)

(7) They found no problem (or a little difficulty) picking clothing from the floor.

(8) They found no problem in household chores

(9) It was impossible to cut their toenails

Use of aids

A walking cane (stick) was used by two thirds of patients and users were more likely to say they used this aid "often" than "sometimes". Crutches and frames were used by less than 10% of patients. In all 70% of patients used at least one of these aids at least some of the time.

Not all patients managed to answer this question as intended and would answer only one item. If answers were obtained for only one or two items then it was assumed that for the unanswered item the device was never used.

Joints replaced

Right knees were a little more likely to have been replaced. Combining the percentages showed that 44% of patients had undergone replacement of both knees, 31% of their right knee alone and 25% their left knee alone. As many as 28% had undergone replacement of another joint other than a knee (usually a hip).

Patients frequently failed to tick a box for all five items and absence of a tick for **some** of the five joints was taken to mean that those joints were not replaced. Sometimes a single item was ticked in the "No" box, and this was taken to mean that this was in fact the only joint which had been replaced.

Joint pain

This question was poorly answered by respondents with no tick given by at least 20%

concerning knee pain and over 40% for other joints. It was not as easy to interpret an isolated tick in the "No" box as for the items on whether joints had been replaced. This was because although some degree of pain may be assumed, it was unclear whether this was while moving or while resting.

Among those who did reply to the various items, over half reported pain in knee joints and between a quarter and a third for other joints. Among the 678 reporting a replaced right knee, the distribution of reported pain was surprisingly similar to those who had not had their right knee replaced. The same was true for the left knee.

Sociodemographic questions

The vast majority of patients were married or widowed with the married being most common. Most patients reported themselves as owner occupiers (61%) or council tenants (24%). Sixty two per cent reported living with either a spouse, child, sibling, other relation or friend. These mostly consisted of the married, although one sixth had a son or daughter living with them.

Knee items

For every one of the seven visual analogue items, at least 40% of responders to the item ringed the best possible option. This was particularly clear for item(5) which asked how worthwhile the knee replacement had been, and at least two-thirds of patients thought it very worthwhile. The items on pain showed more patients prepared to admit some degree of pain on moving than while resting, as expected. In item (3) patients were asked to compare the situation with one year previously and nearly 20% chose the middle option which indicated no change. However 40% rated their knee much better than a year ago.

Half the patients would have liked to bend their replaced knee more and one third could not sit with their knee bent in a confined space for more than 10 minutes.

Distribution of knee score

The distribution of knee score is shown in Figure 10-1. This indicates a negatively skewed distribution. Twenty per cent of patients gave the perfect score of 42, and at the other extreme were 8 patients (1%) with a score of 0. The mean was 32 and the median was 35. The interquartile range spanned 26.5 to 40.5, so that the knee score exceeded 26 in three quarters of patients.

Comments

Positive comments were written by 41% of patients. In all 34% made negative comments, including 11% who spoke of the pain they were still experiencing, 10% concerning the lack of flexion in their new joint and 19% who made a variety of other negative comments. In all 26% of patients made comments about other aspects of their medical history which bore on why the benefits of the new knee joint were not fully experienced.

Demographic variable distribution

Seventy six per cent of the sample were female. Concerning the type of disease requiring knee replacement, 44% had OA, 26% RA, 5% other diagnoses and 25% unknown. These were taken from the surgeons' notes or computer records and were inevitably incomplete.

Concerning age at the time of knee replacement (the average age if both knees had been replaced at different times), 13% were under 55, 23% were 55 to 64 years, 31% were 65 to 74, 13% over 75 and age was unknown for 20% due to not knowing the date of birth or date of operation). As with other basic demographic items, these were more complete for some surgeon series than others.

The time since operation varied up to a maximum of 17 years before the questionnaire was mailed. Again average time was taken for patients who had both knees replaced at different times.Fourteen per cent had received their knee replacement in the last three years, 31% between 3 and 6 years, 32 per cent between 6 and 10 years, 13% over 10 years ago and the data was unknown for 10%.

Distribution of potential confounding variables between series

Distributions of replies to each item according to surgeon series were inspected. Variation by series was seen in the majority of the items concerning daily activities (items contained in the function score), in the use of aids, in the proportion who had both knees replaced, in the pain experienced in joints other than the knees, and in some social factors. In general patients in series 04 seemed to be experiencing greater problems than average with both the pain and the functional capabilities of their joints while patients in series 07 seemed to be least afflicted. Hence any comparison of knee score between series will need to take into account a number of confounding variables.

Knee score and surgeon series

Figure 10-2 shows the distribution by series. The series with the lowest mean and median scores were 03, 04 and 05. Series 07, 08 and 09 came highest in terms of both mean and median. The order differed slightly however; for means it was 07, 09, 08 whereas for medians it was 08, 09, 07. The spread of results was least for series 07, suggesting that this series had consistently good results but fewer "perfect" scores of 42 than series 08 or 09. Series 08 had a larger number of poor results and its distribution showed a longer "tail" at the lower end of the distribution.

The whole analysis was repeated by computing a total score which excluded the use of item k3. The ranking of the ten series was scarcely altered.

Knee score and other factors

There was no particular evidence of difference in the distribution of knee score (using Kruskal Wallis tests) according to gender, type of arthritis, whether one or both knees were replaced, marital status or age group. There was a slight tendency for knee score to decline with age at operation with median score declining progressively from 37 in the under 55 group to 34 in the over 75s (Figure 10-3). A slight tendency for knee score to worsen with time since operation was observed ($p=0.05$, Figure 10-4). No overall

difference was observed according to housing status ($p=0.19$) although the small number in old people's homes expressed less satisfaction (median score 28, see Figure 10-5). Owner occupiers appeared more satisfied than council tenants (medians of 36 and 34 respectively).

There was a highly significant relationship of knee score with use of aids ($p<0.0001$). Median scores were 39.5, 37 and 31.75 for those who never, sometimes and often used aids. Those not answering had a median score of 31. (Figure 10-6)

Reported change over one year by time since operation

Since those operated upon longest ago reported least satisfaction according to the knee score, item K3 was investigated. This item asked subjects to compare their feelings about their knee replacement now compared with one year ago. Figure 10-7 shows that the majority of patients operated upon in the previous three years were particularly enthusiastic about the improvement obtained over the last year. Such improvement was less evident in responses from those operated upon between 3 and 6 years previously, while those receiving their knee replacement more than six years ago seemed even less inclined to report improvement. The median score among this group was close to 3 which indicated no change either for better or worse was perceived.

Comparison of series using the same prosthesis

Using Figure 10-2 to compare series 07 and series 10 suggests that the former did considerably better. However surgeon S7 (pioneer) made an important modification to his prosthesis which is believed to have led to improved results. Surgeon S10 has used the modified prosthesis but has only provided data on the old form of prosthesis. To make a fair comparison between these two series must involve only those patients from series 07 who used the old form.

Sixty four out of 138 left knees and 64 out of 146 right knees were identified as definitely being the old style of prosthesis E in series 07. These were compared with 26 left knees

and 27 right knees in series 10. The results for this subset of series 07 were not as good as for the whole series. Indeed the median total scores for both left and right knees became very similar to series 10. For right knees the mean score was still lower for series 10, but mean scores were similar for left knees.

This suggests that there is no dramatic surgeon effect for prosthesis E after all, and that some of the benefit of series 07 is owing to the new form of the prosthesis and perhaps the patients to whom it is now restricted. It is now restricted to osteoarthritis patients whose anterior cruciate ligament is present and undamaged (Carr et al 1993b).

The different results obtained for series 04 and 06 (same surgeon using prosthesis C) cannot be explained by length of time since surgery, since average time was similar in the two series. It was shown that series 04 consisted of a higher proportion of council tenants. It is possible that surgeon S4 with a split commitment between two hospitals, was better able to procure good rehabilitative care in the specialised orthopaedic hospital (series 06). Both series consisted of a high proportion of rheumatoids, with slightly more in series 06.

Surprisingly similar results were seen for series 01 and 02. These two surgeons , working in hospitals not far from each other, are both using a prosthesis designed in the USA. Surgeon S1 has a longer experience as an orthopaedic consultant and it is possible that surgeon S2 followed the example of S1 in using prosthesis A. If anything series 01 has slightly better results except for degree of flexion, when series S2 is superior.

Odds ratios in unifactor and multifactor analysis

Logistic linear models were fitted to the data on 819 patients for whom the knee score was calculable as a good or poor result. Table 10-2 shows the odds ratios for obtaining a good result according to each factor in turn. This table includes both unadjusted (unifactorial) and adjusted (multifactorial) analysis. In the latter case the influence of each variable was adjusted with respect to the others. Significant relationships were found between the outcome and surgeon, use of aids, and time since operation. No clear

relationships were obvious with arthritis, number of operations (not shown), marital or housing status or age at operation.

The effect of surgeon series had a chi-squared value of 22.5 with 9 degrees of freedom. When allowing for the effects of each of the other factors, this chi-squared value changed very little, never dropping below 20.0 with 9 degrees of freedom. When allowing for the effects of all the other variables, the chi-squared value was 27.8. This shows that the apparent difference between the ten surgeon series cannot be explained on the basis of any of the other eight factors. If anything the effect was amplified, especially by the time since operation factor (since an analysis which allowed for this factor alone had a chi-squared value of 35.3 for surgeon series).

The adjusted analysis shows that series 04 and 06-09 (the series which include patients operated upon before 1980) were ranked higher than for unadjusted analysis. It was generally clear that good results were more likely with operations performed more recently . Thus series which contained higher proportions of patients operated upon more than ten years prior to the survey improved their relative performance when the time factor was allowed for. For series 07,08,09, who already were ranking high in the league table, the rank became even higher. Series 04 and 06 which had been doing badly before (especially series 04) have been partly exonerated.

No other factor than time since operation appeared to modify the difference between surgeon series. This relationship is now shown in Table 10-3. Within nearly every series, knee score declined with time since operation.

Analysis of the four main prostheses

In Chapter 6, it was described how the original plan was to study patients only from Series 04, 06, 07, 08, 09. This was because these were among the oldest and largest series in the UK. Hence multifactorial analysis was repeated on only these series. In addition series 04 and 06 (same prosthesis, same surgeon, see Chapter 7) were combined. This led to a comparison of four prostheses involving 567 patients.

Unifactorial analysis gave the surgeon effect a chi-squared value of 10.9 with 3 degrees of freedom whilst multifactorial analysis increased it to 14.2. In each analysis, series 04/06 was worst and series 09 was best. Series 07 did better than series 08 in unifactorial analysis but less well in multifactorial analysis.

Interactions were sought with gender, diagnosis, age and time since operation. Possible interactions were shown for time since operation and type of arthritis ($0.05 < p < 0.1$ for each). The overall impression of a decline in satisfaction with time since operation held for every series except 06 alone, but this is probably a random finding. Rheumatoid patients were more satisfied in most series except for Series 07 (see Table 10-4). The newer form of Prosthesis E is now reserved chiefly for OA patients, in whom better results are expected. Therefore analysis was carried out on patients who underwent knee replacement with the older form of Prosthesis E. Among these patients, better results were obtained for RA patients (mean score 33.7) than OA patients (mean score 30.4), in line with the trend among the other Series. Therefore the peculiar effect of arthritis in Series 07 is probably due to good results obtained using the newer form of prosthesis E on OA patients alone.

Restricting analysis still further to surgeon series 07, 08 and 09 showed little to choose between the three in overall satisfaction.

Function score

The distribution of this variable is shown in Figure 10-8. The score which in theory and practice varies from a minimum of 0 (worst) to a maximum of 9 (best), is mildly negatively skewed.

The function score varied markedly between the surgeon series, as is shown in the Box and Whisker plot in Figure 10-9 and Table 10-5 . It was greatest for series 07, in accordance with findings on individual function items described earlier in this chapter, and least for series 04. One way analysis of variance showed a very marked difference in mean score ($F(9,797) = 5.40$, $p < 0.001$).

Mean function score was found to be higher in men than women (5.9 vs 4.9), higher in OA patients than RA patients (5.8 vs 4.2), and lower in bilaterally replaced patients (4.8 vs 5.5). Function appeared to increase with age at operation from under 55s (mean 4.8) through 55-64s (5.0) to 65-74s (5.4) but was least for over 75s (4.6). Mean scores decreased with time since operation from 5.9 for those within 3 years post operation to 4.2 for those operated upon more than 10 years ago. Patients not requiring aids obviously had greatest functional score (6.4) compared with those sometimes requiring aids (5.6) and those often requiring aids (4.4). Married people were more functional (mean 5.3) than single people (4.6), with other categories occupying intermediate positions. Owner occupiers had a higher mean functional score (5.4) compared with council tenants (4.7), while those in old people's homes had a mean of 3.3.

A multifactorial analysis which allowed for the effect of each of the other eight factors in turn showed that the surgeon effect was not explained by any of them. The surgeon effect diminished slightly when allowing for type of arthritis ($F(9,794) = 3.61$, $p < 0.001$) but scarcely altered the rankings of the series. Conversely the surgeon effect increased in strength when time since operation was taken into account ($F(9,793) = 7.07$, $p<0.001$). As for the multifactorial analysis of knee score reported above, the 'older' series improved their rankings, series 04 now occupying a medium rather than a low ranking, while series 07 still appeared to have the best mean function.

It is clear that more of the factors considered showed a significant relationship with function score than they did with knee score. This would support the claim that the success of a knee replacement cannot be judged just in terms of function since it clearly related to other factors. However the difference between surgeon series in functional scores has not been explained on the basis of factors measurable in this study.

Comments

The proportion of patients making any comment at all ranged from 61% to 80%.

Positive comments

The difference in proportions of positive comments between surgeon series was not significant (χ^2_9 = 13.5, p=0.15). According to unifactor analysis, they were more frequently made by bilaterally replaced patients, by owner occupiers and by patients under 55. They were made less frequently by patients who often required aids and by single and divorced people. Patients operated upon over 10 years ago were surprisingly most likely to make positive comments, whereas fewest positive comments came from patients between 3 and 10 years post operation. No effect of surgeon series could be demonstrated even after multifactorial analysis.

Negative comments

Comments critical of the degree of bend obtainable from the replaced knee ranged from 4% to 20%, those which mentioned pain as a problem from 8% to 18%, while negative comments other than the two types described above were elicited in 19% of cases, ranging from 15% to 30%.

When all types of negative comments were pooled, 35% of patients made some sort of negative comment, and this was highest for Series 04 (47%) and lowest for Series 07 with 30%. However the proportion of any type of negative comment did not clearly differ between surgeon series (χ^2_9 = 12.4, p=0.2). However there was more of a difference for comments relating to the degree of bend (χ^2_9 = 21.5, p=0.02), and for other negative comments (χ^2_9 = 15.9, p=0.07), but no relationship with comments concerning pain (χ^2_9 = 8.1, p>0.25). Comments about deficiencies of bend were most likely to come from Series 05, while other negative comments came most frequently from Series 04 and 06.

Negative comments were less likely to be made by RA patients, by those not needing aids, by council tenants and by patients aged 55-64. They were more likely to be made by single people, by owner occupiers and by patients operated upon within the last three years.

Some weak evidence of a surgeon effect appeared when type of arthritis was included in a logistic model (χ^2_9 =15.7, p=0.08). Rankings of series now showed that surgeon series 04 and 06 were most likely to elicit negative comments (same prosthesis, same surgeon).

Since certain subgroups of patient seemed most likely to make both positive and negative comments (for example owner occupiers), it must be assumed that the frequency of comment making is culturally determined and seems to tell us little about success of the operation. However the median knee score was 38.75 for those who made positive comments and 33 for the rest. Also negative comment makers had a lower median score of 32 than all others.

Finally it may be noted that patients operated upon within the last three years seemed more likely to make negative comments, despite their relatively high knee score. This suggests their awareness of potential problems of the knee replacement may be higher and they are keener to volunteer information.

Early vs late responders

The knee score was compared between patients who replied quickly to the questionnaire compared with those who responded only after a reminder. The mean score was greater in the early responders (32.4 vs 29.5). The proportion exceeding a score of 35 was 51% and 34% respectively. The direction of this difference held in eight of the ten series.

Discussion

General differences among surgeon series

This study has shown marked differences between surgeon series in patients' degree of satisfaction with their knee replacements. These differences are not explained by other factors with potential influence, namely sex, age, diagnosis, time since operation, number of knees operated upon, marital and housing status. Of course these factors relate to

satisfaction in themselves but their differential distribution among surgeon series does not explain away the apparently superior results in series 07, 08 and 09. In addition, allowing for the effect on satisfaction of aids used does not make any difference either.

One of the more curious findings is that the magnitude of the difference between series is amplified when allowing for the length of time since the operation. Surgeon series 07, 08 and 09, who already were shown to have superior results were in a sense handicapped by having a higher proportion of patients operated upon in the 1970s. These patients had a lower knee satisfaction score. Therefore the generally good results for series 07, 08 and 09 were magnified by multifactorial analysis. Similarly the poor results for series 04 and 06 were mitigated in multifactorial analysis. The rankings of the remaining series, all of whom only had patients from the 1980s, dropped in multifactorial analysis.

It would therefore appear that the older surgeon series have obtained particularly high accolades from patients operated upon more recently. There are two possible explanations for this:

1. The success of knee replacements increases with the experience of the surgeon. It is possible that the surgeons responsible for the 'younger' series have not reached the peak of their learning curve. However the learning curve for operative skill is unlikely to continue for a full decade. All prostheses from the older series underwent modifications and these may have been the cause of increased success.

2. Patient satisfaction with a knee replacement decreases with time since operation.

The evidence from item K3 of the questionnaire was that patients perceived little improvement in their knee function after 6 years, whereas they perceived great improvement in the first three postoperative years.

However a cross sectional study such as this cannot distinguish between these two possible explanations. If this questionnaire was sent in 1994 (say) to those patients in the

present study who were operated upon between 1987 and 1990, evidence for or against the second explanation would be provided.

Differences between the older surgeon series

Series 04 and 06 have in every outcome measure demonstrated consistently poorer results than series 07, 08, and 09. Curiously series 04 has done worse than series 06. Prosthesis C, which was used for both these series, is of a variety which was condemned recently (Noble and Hilton 1991). It is a fully constrained hinge prosthesis which does not allow rotation and is liable to react badly to sideways force. It is now seen by orthopaedic surgeons as an option to be reserved when less constrained prostheses have failed (Lancet 1986b, Noble 1990), and this study would support such a policy. Survival rates for hinged prostheses were shown to be lowest among both OA and RA Swedish patients (Knutson et al 1986), and deep infection was commonest amongst hinged prostheses in a retrospective review of knee replacements carried out in Bristol (Johnson and Bannister 1986). The large American study of Rand and Ilstrup (1991) suggested that the "older constrained" prostheses had a slightly lower 10 year survival rate than the average for the total series. However the comparison by Tew et al (1985) of failure rates for nine prostheses suggested that hinged prostheses performed neither worse nor better than the average.

Some association may be detected between the results of the various series with respect to knee score and the response rate to the questionnaire (see Chapter 8). Among the older Series, 04 and 06 had poorest response while Series 07 and 08 had a relatively good response. However Series 09 with its good results did not have a good response. There was evidence that patients who only responded after a reminder were less satisfied than the early responders. Perhaps non respondents would have been even less satisfied. This implies that the relatively poor results of Series 04 and 06 are likely if anything to be underestimated by this survey. Series 09 may have obtained slightly flattering results but the relative success of Series 07 and 08 is likely to be genuine.

It was clear that series 07 included patients whose degree of arthritis was on average less

marked. The surgeon responsible has narrowed his criteria over the years and now only uses a unicompartmental type, reserved chiefly for OA patients. It was interesting that better results were achieved for OA patients in this series alone and this was due to particularly good results in OA patients using the unicompartmental prosthesis.

The comparison of series 07 with series 10 (prosthesis E) which included patients with only the bicompartmental type showed very little difference. This is encouraging since it shows good results may be reproduced by surgeons other than the prosthesis pioneer. If anything series 07 lost ground to series 08 and 09 in multifactor analysis (the former being penalised for the apparently stricter selection criteria).

A comparison of the results for prosthesis D done in series 05 (surgeon not the inventor) and series 08 (surgeon the inventor) revealed a striking difference in results obtained. There were no major differences obvious in the characteristics of patients in the two series, except that series 08 were more likely to consist of widow(er)s while series 05 contained more single people. Series 05 were more likely to be home owners. However social characteristics made little apparent difference to outcome in this study. There were no differences in the proportion of patients having experienced hip replacement or recent pain in hips or ankle joints. The other difference is that series 05 started around 1981 while series 08 did so in 1976. The difference in mean time since operation was marked but but the apparent effect of this factor shown above would if anything amplify the superior results of Series 08. Surgeon S5 (responsible for series 05) was aware of suboptimal results with regard to flexion obtainable with prosthesis D, and in September 1987 began operating with a slightly modified form of prosthesis D. However since surgeon S8 never made any such modification yet managed to procure superior results, prosthesis design cannot be the only explanation.

The apparent surgeon effect for Prosthesis D is surprising since it has been claimed to be easy to insert in the manufacturer's promotional literature. Prosthesis D, though not fully constrained, includes a long femoral stem, and by the admission of surgeon S8 (pioneer) cannot produce more than 90 degrees of flexion. In general it would be seen as belonging to an older generation of prostheses. This study has underlined the

possibility of obtaining excellent results but equally the possibility of poor results.

Series 09 has attained the top ranking according to this study. Surgeon S9 is highly experienced and respected but his prosthesis has undergone a number of modifications over the last two decades. Results in the literature presented for this prosthesis have distinguished between early and late marks of the prosthesis. This study has included too small a sample size for comparing results within series 09. A fairly low response rate was obtained (probably because this is the oldest series) and these were exclusively from surgeon S9's NHS patients which comprised only two thirds of his full series. It is possible that more articulate private patients would have had higher expectations of surgery and would have voiced more criticism. However this series has not shown less satisfaction among owner occupiers than council tenants, in fact if anything they appeared more satisfied. The 'social class effect', if it exists, is unlikely to upset the findings for series 09. Prosthesis F seems to be currently respected by the European orthopaedic community, falling well within the mainstream of modern minimally constrained prostheses. This study did not include a further series which uses it, and even Series 09 finally comprised a small sample. Thus the seemingly optimistic results must be regarded circumspectly.

It would however appear that prostheses akin to the current version of prosthesis F, and the unicompartmental version of prosthesis E, provide the best options available. A fuller prospective evaluation is however required and this will be discussed in Chapter 12.

Table 10-1. Distribution of Total Knee Score by Surgeon Series

Series	n	Mean	S.D.	Median	Interquartile range
01	50	31.9	10.8	36.25	27.75 - 40.25
02	49	31.3	10.9	33.0	24.0 - 42.0
03	30	27.2	11.9	29.5	20.9 - 37.6
04	49	29.5	10.6	32.5	21.75 - 38.5
05	83	29.8	10.7	32.5	23.0 - 39.0
06	86	31.6	10.0	34.25	27.25 - 40.0
07	199	33.7	9.2	36.0	31.0 - 40.5
08	169	32.5	11.1	37.0	27.0 - 42.0
09	65	33.2	9.6	36.5	29.0 - 40.0
10	39	31.2	12.2	36.0	23.5 - 42.0

Table 10-2. Estimated Odds ratios (and 95% Confidence Interval) obtaining Knee score of 35 or less or result compared with subgroup first-named.

	Unifactorial analysis		Multifactorial analysis	
FACTOR	Odds ratio	95% confidence interval	Odds ratio	95% confidence interval
Surgeon Series				
01	1.0		1.0	
02	1.05	0.51-2.58	1.34	0.63-2.81
03	2.00	0.77-5.18	1.42	0.58-3.43
04	2.04	0.91-4.54	1.13	0.51-2.48
05	1.54	0.74-3.23	1.86	0.93-3.70
06	1.49	0.73-3.05	1.07	0.53-2.19
07	0.74	0.39-1.42	0.47	0.24-0.93
08	0.85	0.44-1.65	0.43	0.21-0.88
09	0.72	0.34-1.52	0.36	0.17-0.78
10	0.81	0.34-1.92	0.44	0.19-1.01
Gender				
Male	1.0		1.0	
Female	1.23	0.89-1.70	1.27	0.93-1.73

Arthritis type				
O.A.	1.0		1.0	
R.A.	0.88	0.63-1.23	0.72	0.51-1.01
Other	0.85	0.43-1.69	1.03	0.54-1.94
Not known	1.08	0.76-1.52	0.71	0.42-1.17
Aids used				
Never	1.0		1.0	
Sometimes	1.25	0.82-1.89	1.24	0.86-1.79
Often	3.12	2.17-4.48	3.07	2.22-4.24
Not known	3.20	1.82-5.61	3.03	1.83-5.00
Marital status				
Married	1.0		1.0	
Single	0.87	0.52-1.49	0.52	0.32-0.85
Widowed	1.09	0.80-1.47	0.75	0.55-1.02
Divorced/ separated	1.47	0.72-3.02	1.22	0.63-2.36
Not known	1.41	0.58-3.41	1.20	0.35-4.08

Housing				
Council rented	1.0		1.0	
Private rented	0.83	0.43-1.59	0.89	0.50-1.60
Housing Assoc	0.77	0.37-1.59	0.70	0.37-1.35
Owner Occupier	0.77	0.55-1.08	0.79	0.58-1.07
Old people's home	1.83	0.71-4.71	1.65	0.71-3.84
Not known	1.03	0.49-2.18	0.66	0.24-1.78
Time since operation				
Under 3 years	1.0		1.0	
3 - 6 years	1.64	1.05-2.56	1.69	1.07-2.66
6 - 10 years	2.05	1.31-3.20	2.82	1.70-4.67
Over 10 years	1.81	1.07-3.07	2.70	1.51-4.82
Not known	1.94	1.10-3.43	2.49	1.28-4.83
Age				
Under 55	1.0		1.0	
55 - 64	1.45	0.93-2.28	1.37	0.90-2.0
65 - 74	1.51	0.97-2.34	1.25	0.79-1.98
75 +	1.64	0.97-2.78	1.27	0.73-2.22
Not known	1.05	0.66-1.68	0.73	0.39-1.35

Table 10-3. Shows relationship between mean knee score (SD) (n=) and time since operation, according to surgeon series.

	Time since operation (yrs.)				
Series	0-	3-	6-	10-	Not known
01	32.5 (11.9) (n=25)	31.6 (11.0) (n=15)	29.8 (4.5) (n=3)	-	31.0 (9.8) (n=7)
02	33.2 (11.1) (n=26)	29.8 (10.1) (n=22)	-	-	14.0 (-) (n=1)
03	42.0 (-) (n=1)	25.4 (11.6) (n=22)	34.8 (11.1) (n=3)	-	27.8 (13.9) (n=4)
04	-	33.7 (7.0) (n=16)	28.3 (10.3) (n=20)	23.6 (13.1) (n=10)	34.5 (13.0) (n=3)
05	32.9 (9.1) (n=29)	28.6 (10.1) (n=19)	28.1 (8.0) (n=4)	-	27.8 (12.3) (n=31)
06	40.0 (-) (n=1)	29.8 (9.8) (n=36)	36.1 (9.6) (n=17)	31.7 (9.6) (n=19)	29.8 (11.2) (n=13)
07	36.1 (6.7) (n=30)	35.4 (9.3) (n=59)	31.6 (9.8) (n=78)	32.2 (9.3) (n=21)	36.7 (6.4) (n=11)

Table 10-3. Shows relationship between mean knee score (SD) (n=) and time since operation, according to surgeon series.

	Time since operation (yrs.)				
08	36.5 (0.7) (n=2)	33.7 (10.8) (n=55)	31.4 (11.7) (n=82)	33.0 (9.9) (n=30)	-
09	38.4 (2.9) (n=4)	35.0 (59) (n=11)	32.5 (8.8) (n=28)	31.8 (12.7) (n=21)	40.0 (-) (n=1)
10	42.0 (-) (n=1)	38.6 (7.4) (n=6)	28.9 (13.6) (n=24)	-	31.3 (8.6) (n=8)

Table 10-4. Shows relationship between mean knee score (SD) (n=) and type of arthritis by surgeon series.

	Type of arthritis			
Series	OA	RA	Other	Not known
01	32.7 (12.8) (n=7)	32.9 (7.9) (n=4)	-	31.6 (11.0) (n=39)
02	30.8 (12.4) (n=2)	26.0 (-) (n=1)	-	31.4 (11.1) (n=46)
03	-	-	-	27.2 (11.9) (n=30)
04	28.0 (12.4) (n=22)	29.4 (8.4) (n=22)	-	36.6 (9.6) (n=5)
05	28.8 (11.9) (n=21)	30.4 (9.2) (n=14)	-	30.0 (10.7) (n=48)
06	28.5 (11.2) (n=28)	34.1 (8.4) (n=42)	-	30.4 (10.8) (n=16)
07	33.8 (8.8) (n=132)	33.1 (9.8) (n=45)	33.2 (13.5) (n=11)	36.7 (6.4) (n=11)

Table 10-4. Shows relationship between mean knee score (SD) (n=) and type of arthritis by surgeon series.

	Type of arthritis			
08	32.6 (11.7) (n=96)	32.8 (9.9) (n=49)	31.6 (10.8) (n=24)	-
09	31.2 (11.5) (n=35)	35.3 (6.3) (n=27)	36.5 (7.8) (n=2)	40.0 (-) (n=1)
10	28.2 (15.4) (n=18)	35.2 (6.0) (n=9)	33.0 (-) (n=1)	32.6 (9.8) (n=11)

Table 10-5. Relationship between mean knee score and speed of response to questionnaire by surgeon series

Series	Early responders	Reminded
01	34.3	27.2
	(10.7)	(9.8)
	(n=33)	(n=17)
02	31.0	33.1
	(11.1)	(9.8)
	(n=42)	(n=7)
03	27.1	27.9
	(12.7)	(5.7)
	(n=26)	(n=4)
04	30.6	25.8
	(10.8)	(9.5)
	(n=38)	(n=11)
05	29.9	29.1
	(10.9)	(9.8)
	(n=69)	(n=14)
06	31.9	29.2
	(9.7)	(13.0)
	(n=77)	(n=9)
07	34.1	30.7
	(9.2)	(8.9)
	(n=179)	(n=20)
08	33.0	30.3
	(10.4)	(13.4)
	(n=138)	(n=31)

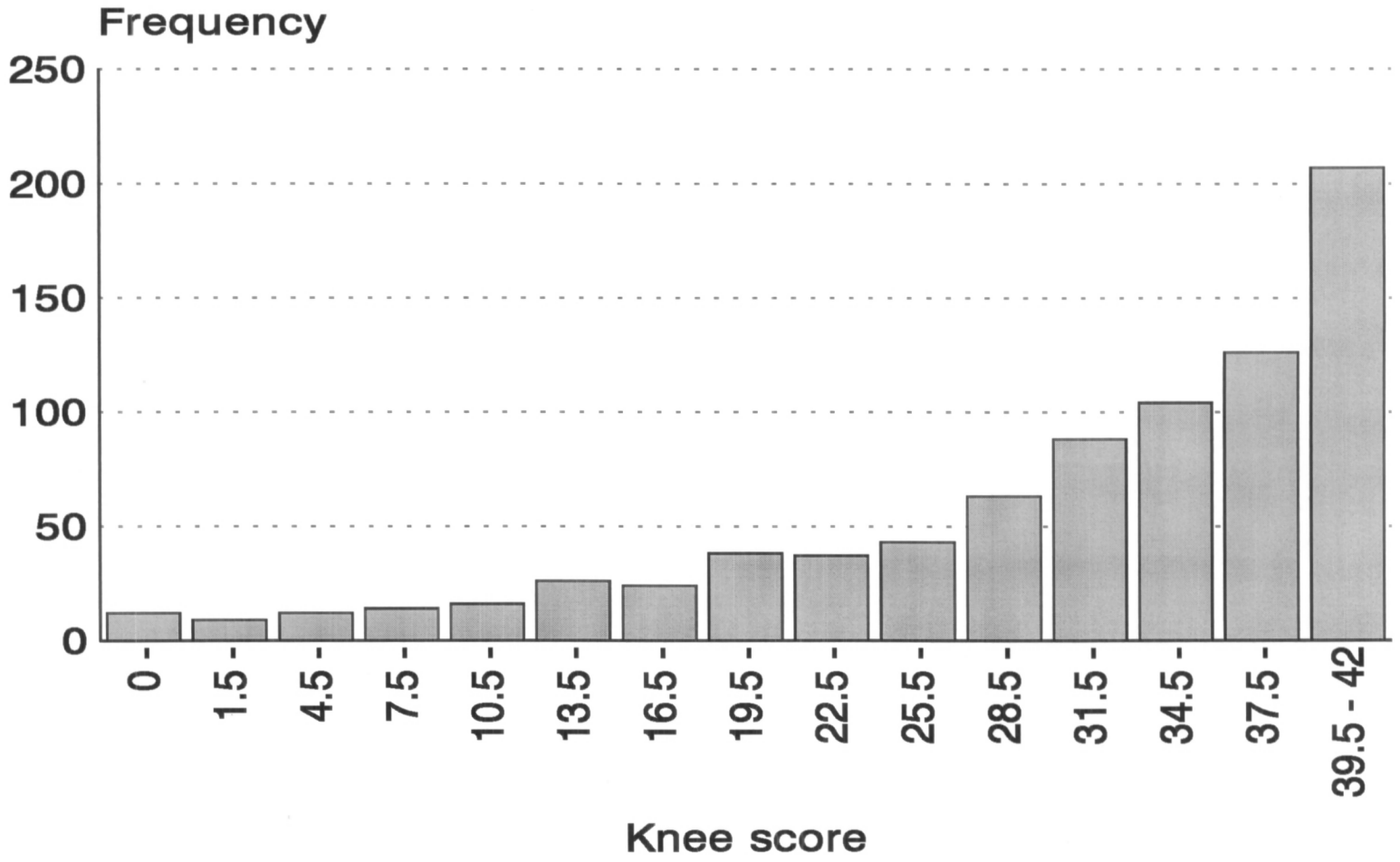
Table 10-5. Relationship between mean knee score and speed of response to questionnaire by surgeon series

Series	Early responders	Reminded
09	33.2	31.0
	(9.7)	(-)
	(n=64)	(n=1)
10	31.0	32.1
	(12.7)	(8.3)
	(n=34)	(n=5)

Table 10-6. Distribution of function score according to surgeon series

Series	n	Mean	S.D.	Median	Interquartile range
01	45	5.4	1.9	5.1	4.3 to 6.9
02	44	5.0	2.6	4.5	2.5 to 7.4
03	25	4.2	2.2	4.3	2.9 to 6.0
04	56	4.3	2.3	4.3	2.8 to 5.5
05	82	4.6	2.0	4.7	1.9 to 6.2
06	86	4.7	2.1	4.9	3.0 to 6.3
07	196	5.9	2.3	6.5	4.4 to 7.8
08	164	5.3	2.3	5.5	3.6 to 7.4
09	71	4.8	2.5	5.0	3.2 to 6.9
10	38	5.3	2.0	5.2	4.0 to 6.5
Total	807	5.2	2.3	5.3	3.6 to 7.1

Figure 10-1. Histogram to show distribution of knee score



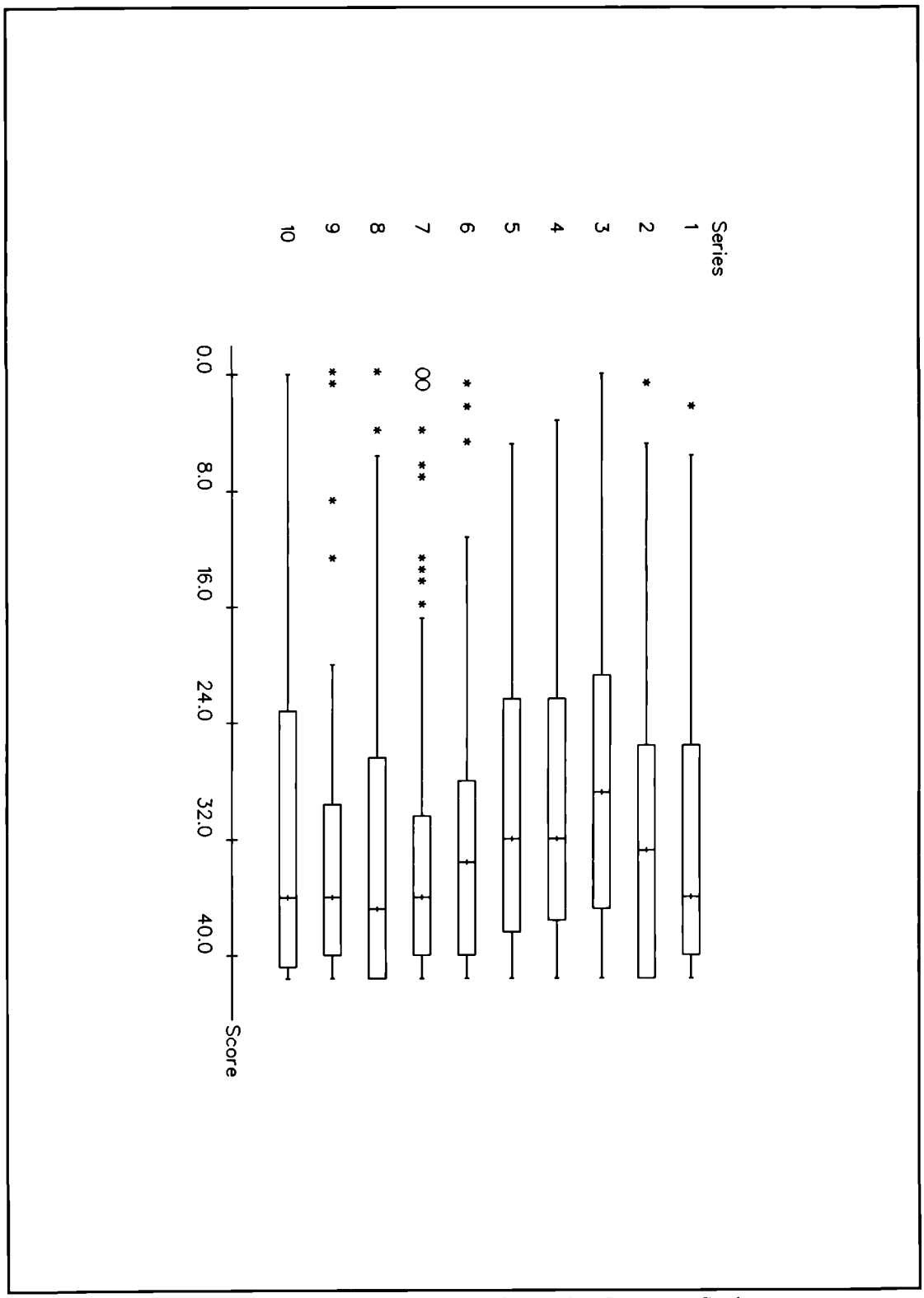


Figure 10-2. Box and Whisker plot of Knee Score by Surgeon Series.

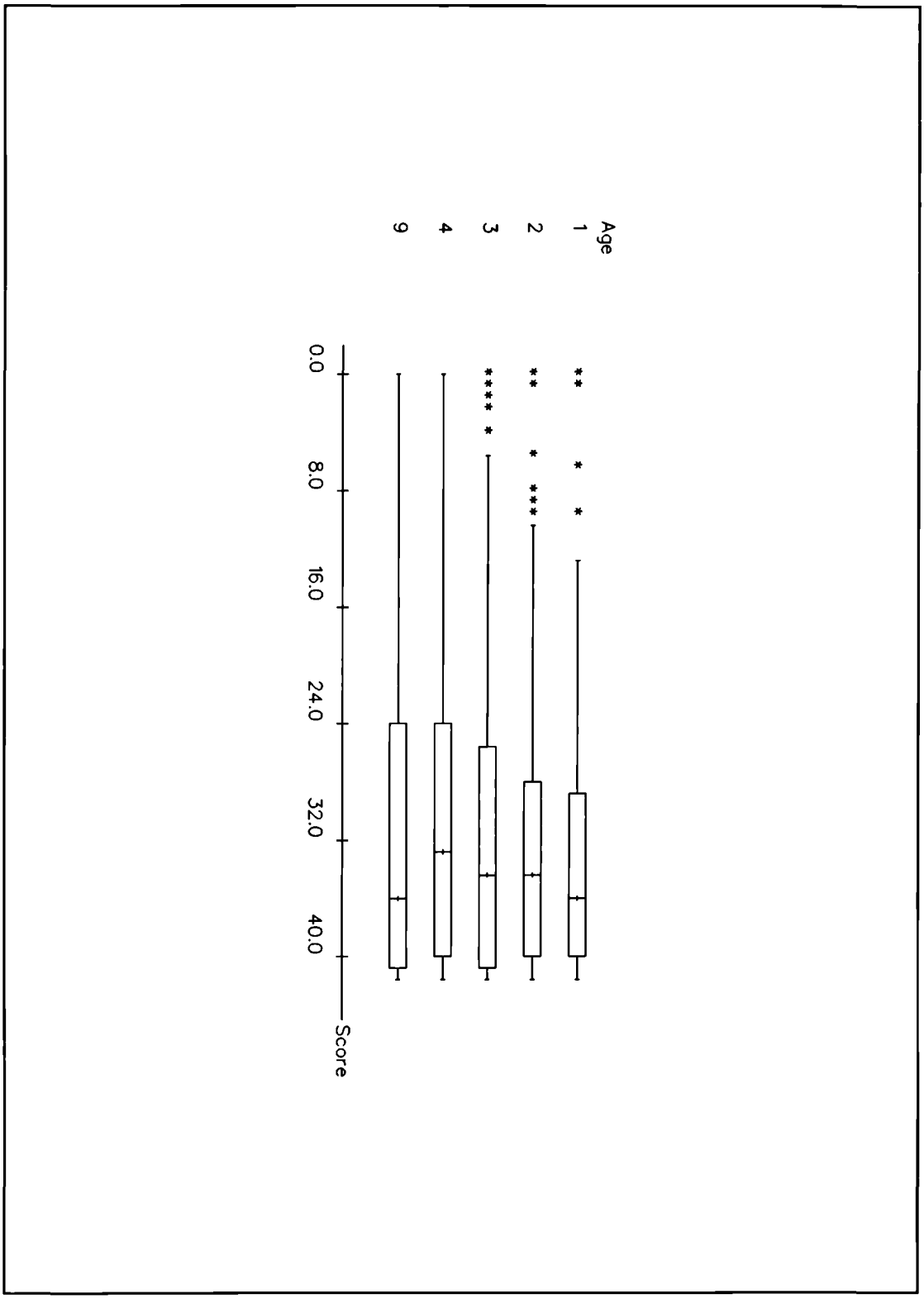


Figure 10-3. Box and Whisker plot of Knee Score by Age Group at operation. (1=Under 55, 2=55-64, 3=65-74, 4=Over 75, 9=Not Known)

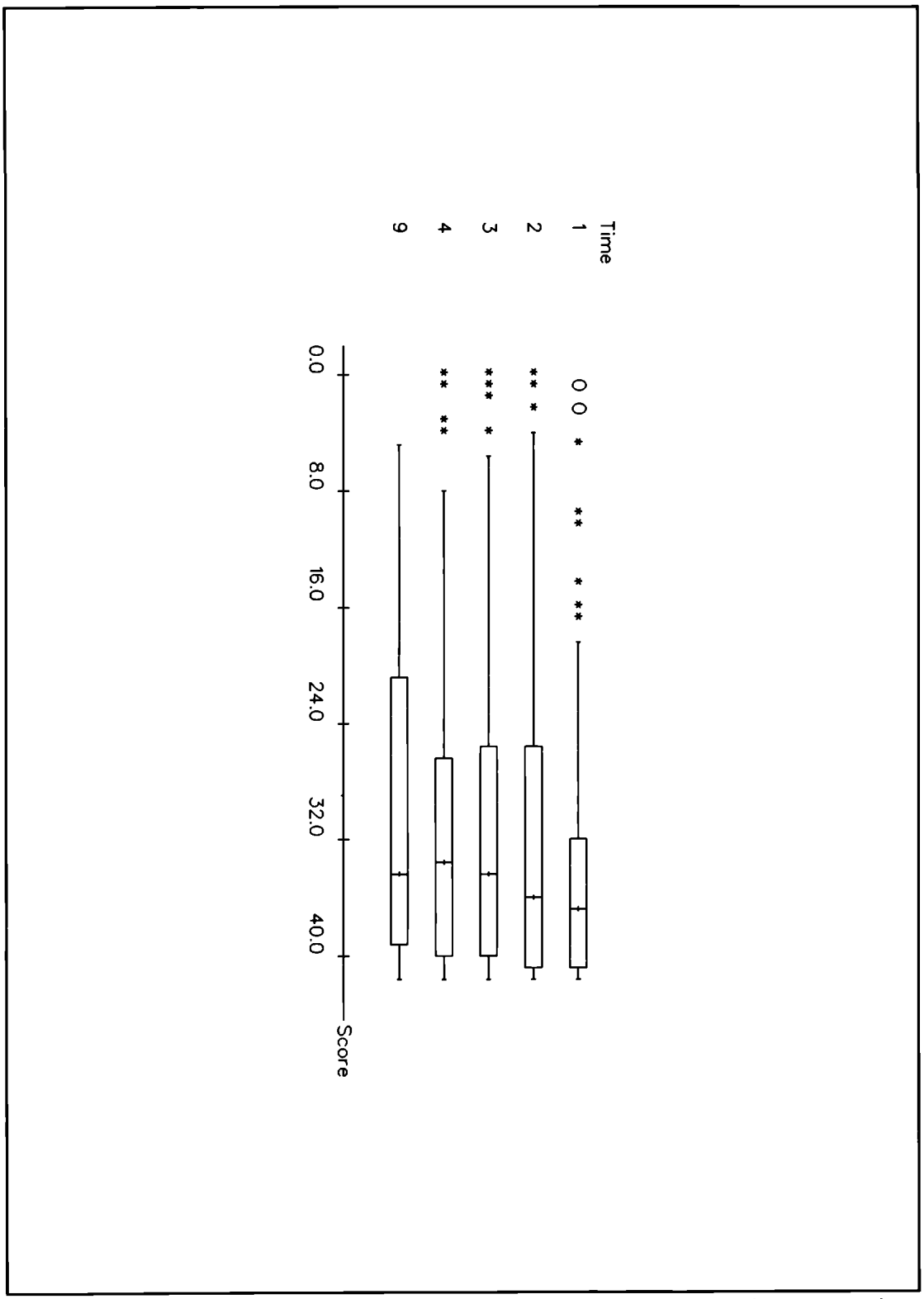


Figure 10-4. Box and Whisker plot of Knee Score by Time since operation. (1=Under 3 yrs, 2=3-5 yrs, 3=6-9 years, 4=Over 10 years, 5=Not Known)

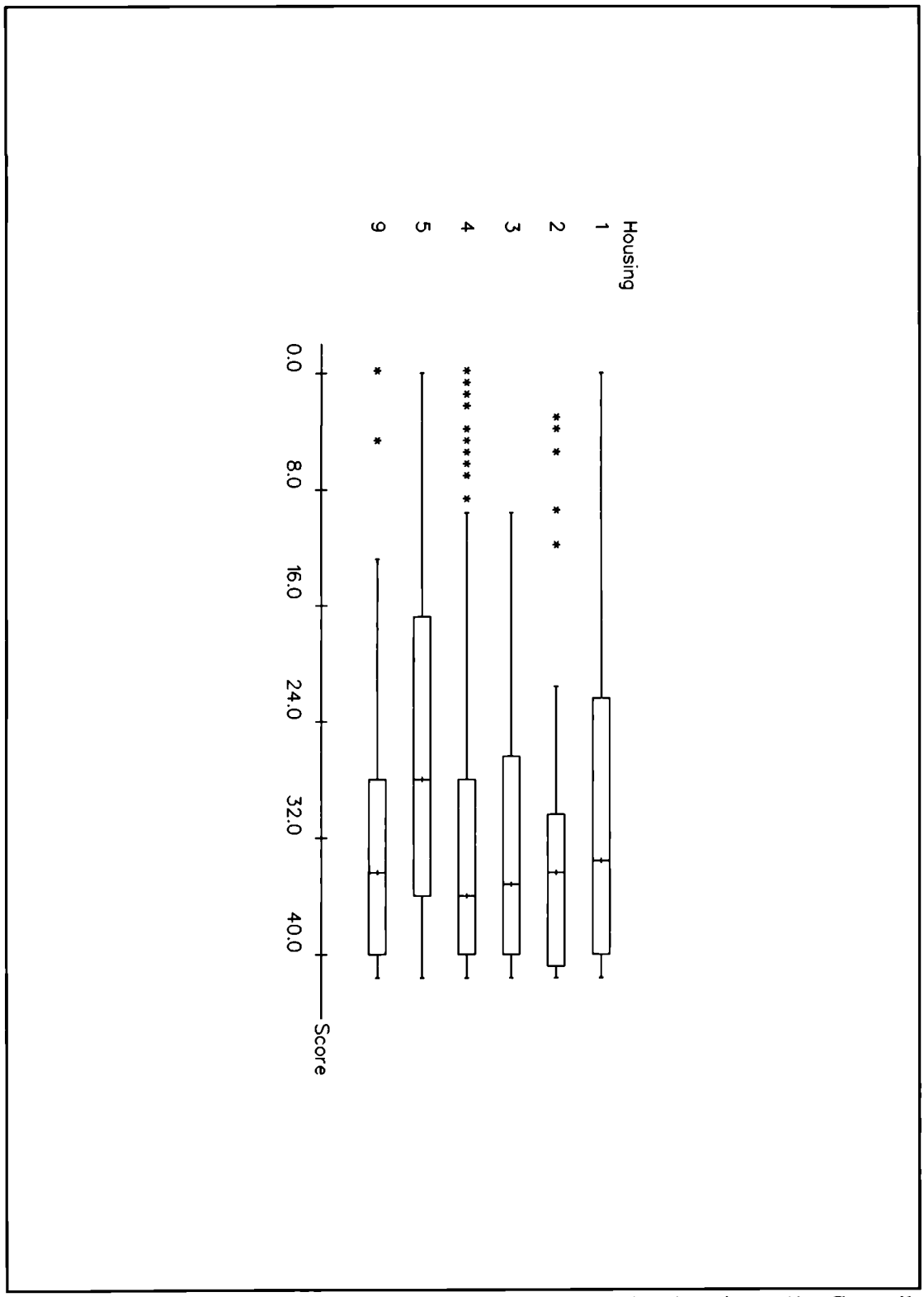


Figure 10-5. Box and Whisker plot of Knee Score by housing. (1=Council, 2=Private rented, 3=Housing Association, 4=Owner Occupied, 5=Old people's home, 9=Not Known)

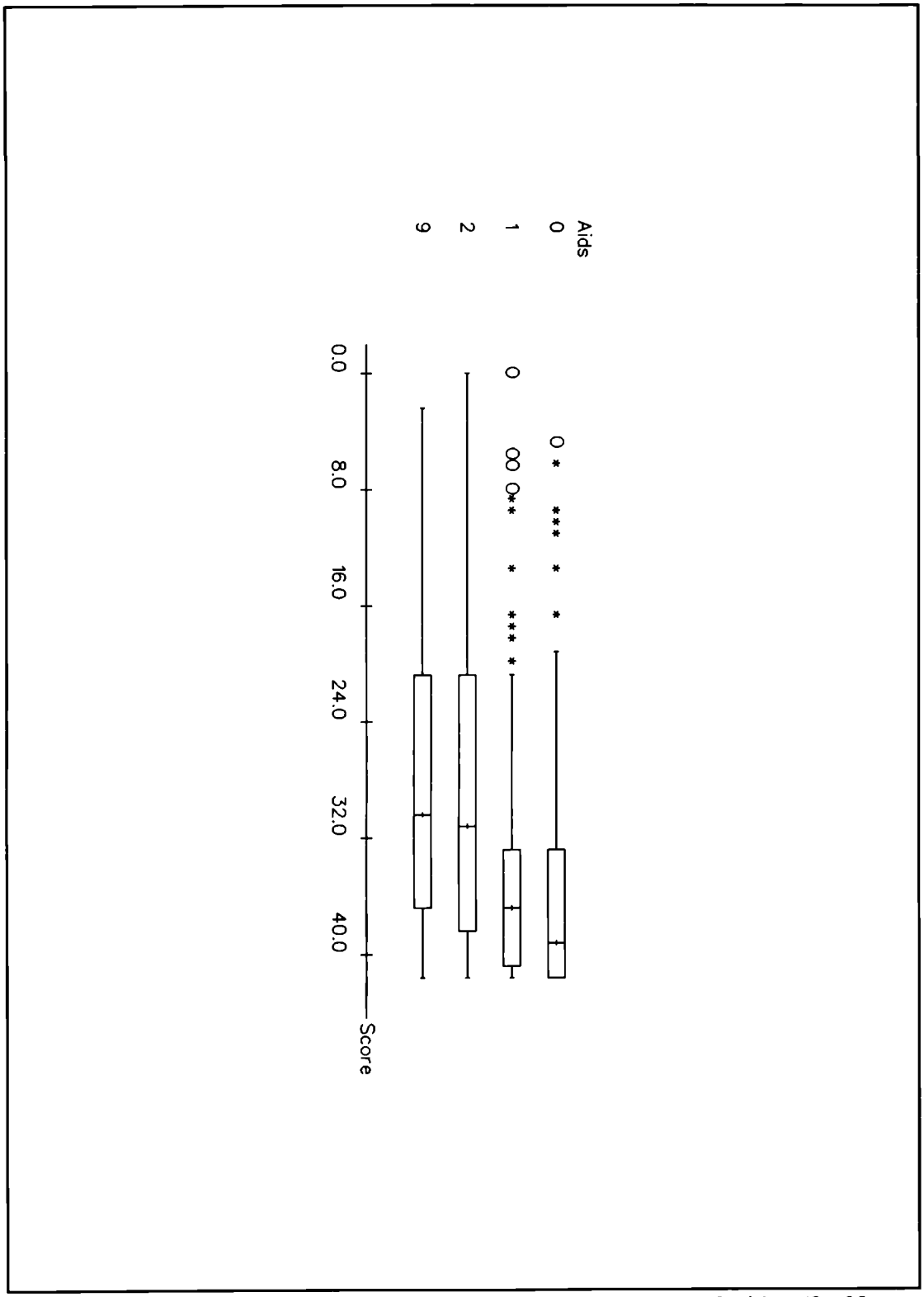


Figure 10-6. Box and Whisker plot of Knee Score by use of aids. (0=Never, 1=Sometimes, 2=Often, 9=Not Known)

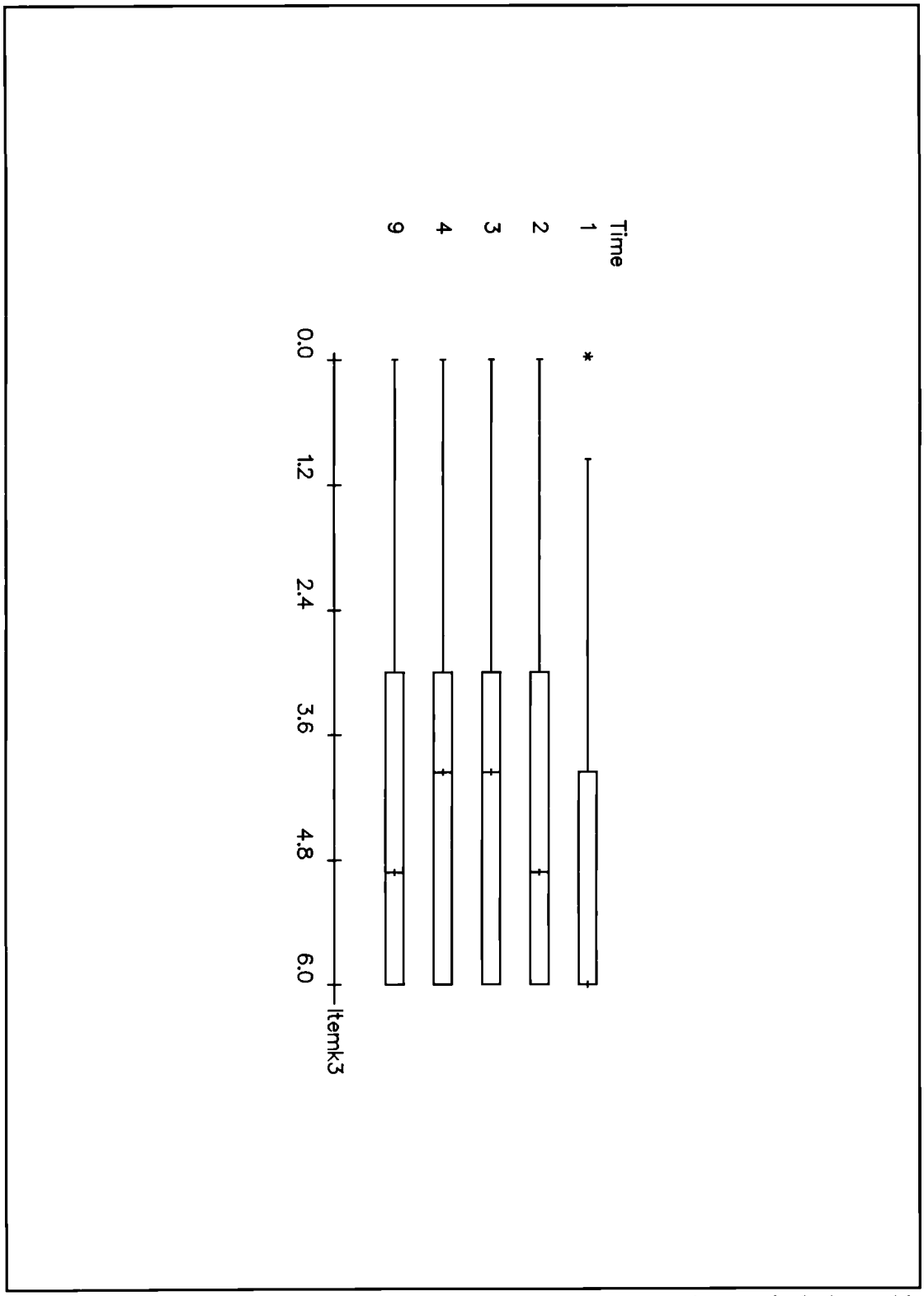
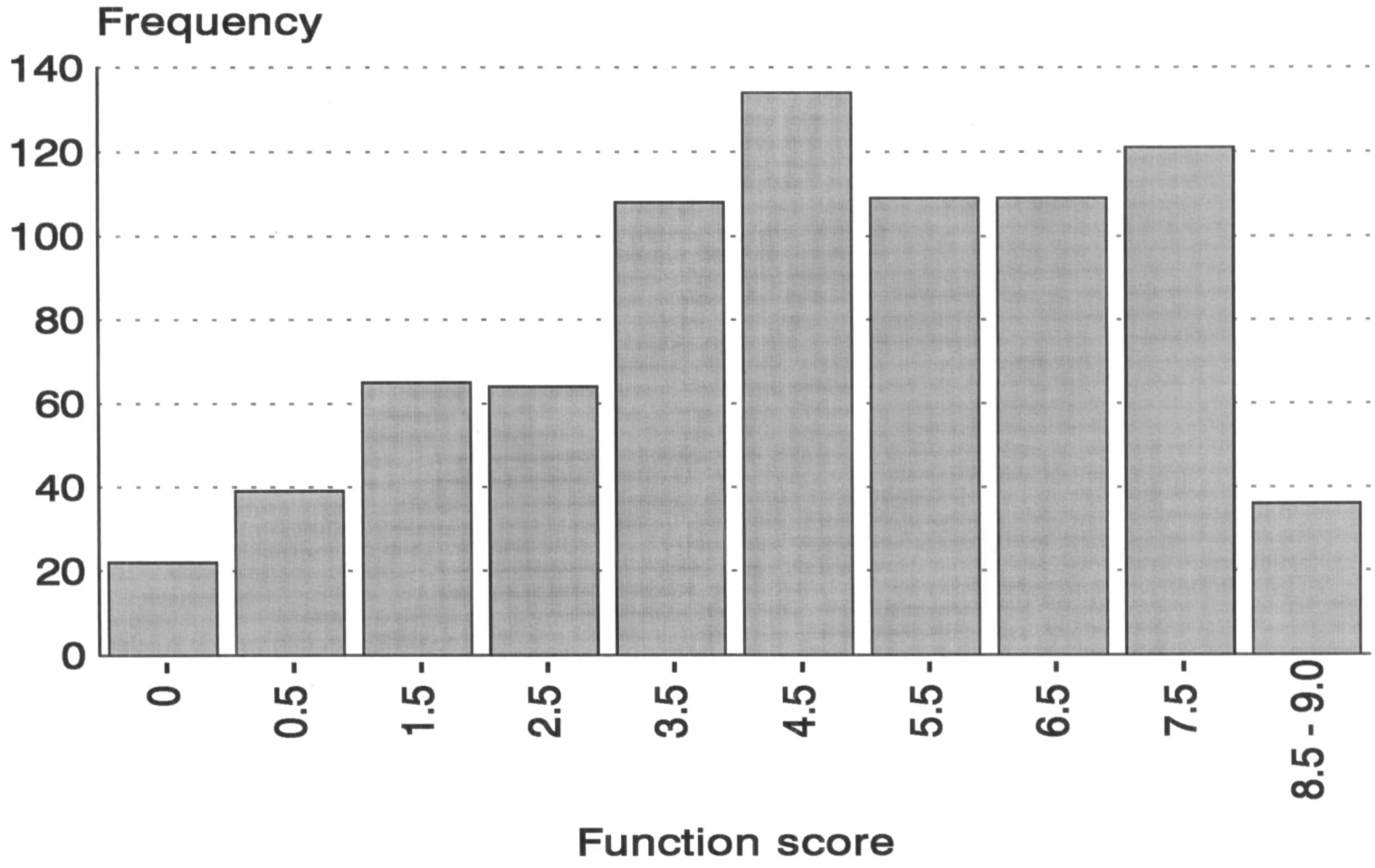


Figure 10-7. Box and Whisker plot of score in item k3 "How do you feel about this knee compared with one year ago?" (0=much worse, 6=much better) by time since operation. (1=Under 3 yrs, 2=3-5 yrs, 3=6-9 yrs, 4=Over 10 yrs, 5=Not Known)

Figure 10-8. Histogram to show distribution of function score



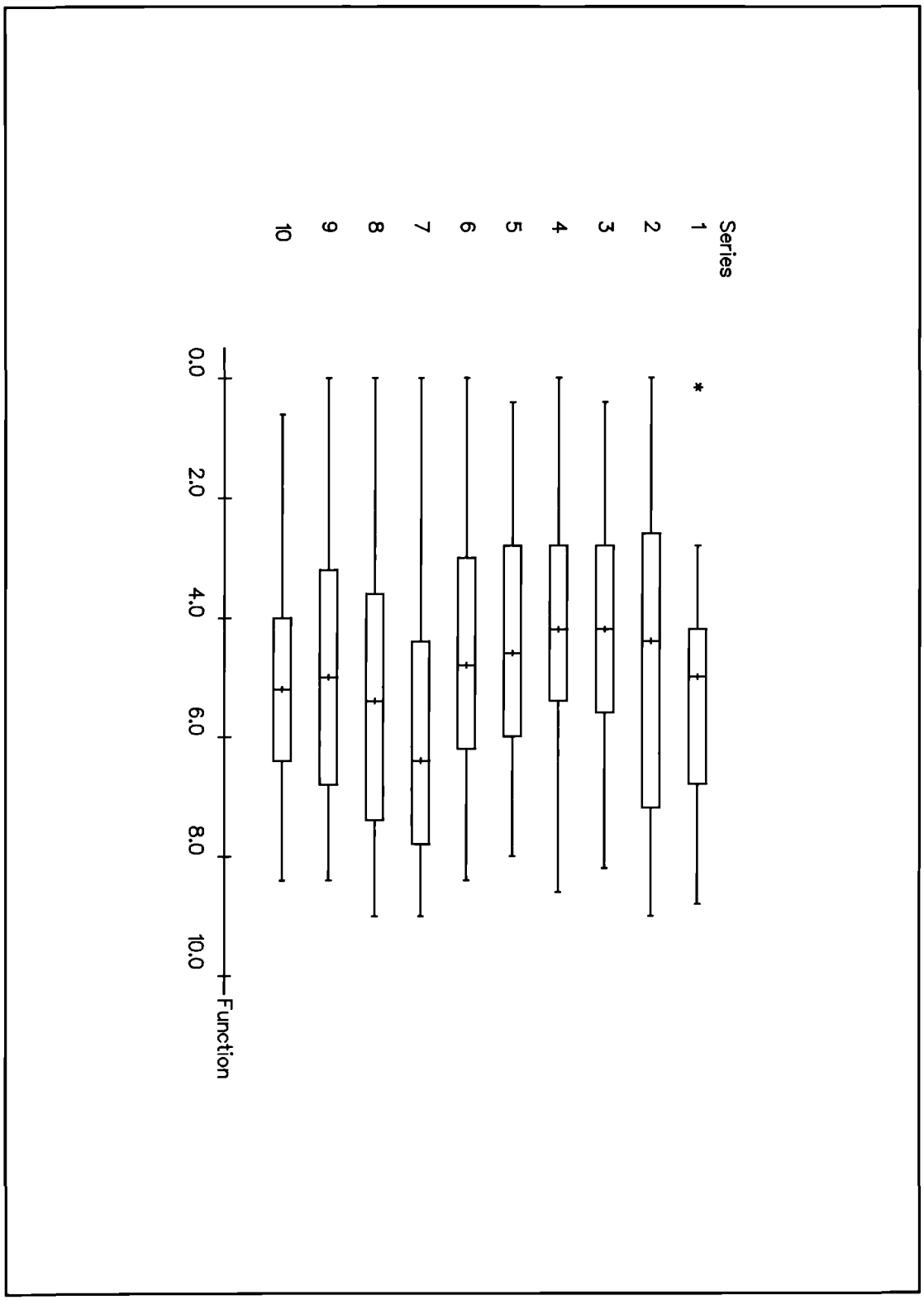


Figure 10-9. Box and Whisker plot of Function Score by Surgeon Series.

CHAPTER 11. COMPARISON OF PATIENT'S AND SURGEON'S OPINIONS

Introduction

Nearly all papers surveyed in the knee replacement literature (see Chapter 3) involved surgeon based assessments of outcome. Sometimes these consisted of radiological measurements and sometimes of clinical assessments of joint movement or functional ability. Frequently points based scores were used based on pain, function, movement and deformity, but these are assessed in outpatient clinics based on the surgeon's judgement. The difficulty of allowing for the gratitude factor has been acknowledged (Thacker and Fulford 1986). As long ago as 1969 one surgeon visited every patient at home to carry out the assessment because of this perceived problem (Bellamy and Campbell 1989). Fitzpatrick (1990) has drawn attention to 'normative effects' which showed among patients in a neurological clinic who felt unable to express dissatisfaction about doctors or the NHS.

More recently Brewster and Newman (1991) reported a comparison of clinic based assessment with patients' responses to a questionnaire administered in clinic. They found a disappointing lack of agreement with surgeons being generally more optimistic than patients. Because of the discrepancy they suggested that postal questionnaires had little part to play in audit of knee replacements. This study together with those of Carlson and Pelletieri (1989) and Black et al (1991) were reviewed in Chapter 4.

It was therefore decided to compare the surgeon's most recent assessments with the replies to the patient questionnaire in this study. However unlike the three studies cited above there was no planning in the timing of the surgeon's assessment relative to the mailing of the questionnaire and it was only possible to make comparisons when patients had been assessed by surgeons in the recent past. Thus the sample size was limited. In addition each surgeon had designed his proforma in different ways so that few items corresponded exactly between the patient questionnaire and surgeon's assessment.

Methods

1. Series 04/06

As outlined in Chapter 5, these two series were operated on by the same surgeon using the same prosthesis in two different hospitals. Patients were operated upon between 1973 and 1988. From 1982, the surgeon was part of the EULAR study and therefore was contracted to carry out regular postoperative assessments at 6 months, 1 year and yearly intervals thereafter. The actual assessments were carried out by the surgeon or one of his registrars. The coordinator of the EULAR study undertakes to provide for surgeons listed data on each operation in terms of the most recent postoperative assessment. Altogether this consisted of 84 knee operations from series 04 and 150 from series 06. Among these were 21 operations (16 patients) in series 04 and 35 operations (24 patients) in series 06 where assessments had been performed in 1990 or 1991. Thus in all 40 patients had recent clinical assessments with which replies to the questionnaire (mailed in the Autumn of 1990) could be compared.

Of the 40, only 32 provided questionnaire data. Of the many items required to be filled in by surgeons for the EULAR study, some bore no relationship to questionnaire items (for example radiological signs), some bore a strong relationship to one or more questionnaire items but the questions and possible answers were not phrased in the same way, and a few items were exactly the same. The items used for comparison were as follows:

1. Standing up from a straight chair

2. Climbing stairs

3. Descending stairs

4. Walking distance

5. Use of aids

6. Pain experienced in operated knee

7. Consideration of worthwhileness of operation

2. Series 05

This consisted of 146 patients operated upon between 1980 and 1989 by a single consultant surgeon. During the months of May to August 1990, the consultant's research registrar visited as many of these patients as possible at home. In the event he managed to see 76, and 64 of these replied to the postal questionnaire, of which 61 replies were considered usable. The postal questionnaire was mailed in August and September 1990 to this series.

The items compared were as follows:

1. Rising from chair

2. Climbing stairs

3. Descending stairs

4. Walking distance

5. Use of aids

6. Pain at rest and during activity

Series 07

Of the 364 patients included in the data base kept for this series by the consultant's academic colleague, 104 had data available both from the postal questionnaire **and** on a recent assessment by the surgeon or one of his registrars in 1989 or 1990. The questionnaire was mailed in November and December 1990.

The items compared were as follows:

1. Rising from chair

2. Climbing stairs

3. Walking distance

4. Pain during activity

5. Consideration of worthwhileness of operation

Series 08

Of the 355 patients (440 knees) included in the series, there were only 50 knees where follow up was known to have been carried out in 1989 or 1990 **and** a postal questionnaire was available. This was partly because the surgeon's data base usually did not contain the actual date of follow up, although it is likely that follow up was in fact carried out in 1989/90 for a substantial number of these patients. Surgeon S8 has in fact carried out all his follow up assessments personally, unlike other surgeons.

The assessments carried out by this surgeon do not involve the same routine questions as for other series. The only items considered compatible with items on the questionnaire

were:

1. Pain at rest and in activity

2. Consideration of worthwhileness of operation (patient)

3. Doctor's assessment

Questionnaire items

Questionnaire items 2-5 (chair rising, going up and downstairs, and walking distance) could be compared fairly straightforwardly with clinical assessments. Use of aids (items 10a-c) was restricted to comparing reported use of walking "cane" (stick). Concerning perceived pain, items 6a,6b of the visual analogue knee oriented questions were comparable with the more qualitative clinical assessment. Finally items 1,2,4 and 5 of the knee questions formed an indirect basis for comparing general satisfaction with the operation.

In order to summarise the results of cross tabulations, trios of "agreement frequencies" are reported. These represent the number of patients who agreed with the surgeon, the number where the patients were more optimistic and the number where the surgeons were more optimistic. For a number of items the definition of agreement is arbitrary so summary statistics must be taken alongside the tables of raw data. In each table of raw data, the asterisk sign shown in some cells of the table is used to show the definition of agreement between surgeon and patient. Tables of raw data for some of the key findings are shown at the end of this chapter while other tables are shown in Appendix 9.

McNemar's test (with continuity correction) was used to assess whether disagreements tended to occur in one particular direction. Two tailed probabilities were calculated against the null hypothesis that disagreements in either direction were equally likely. A number of tables use the knee rather than the patient as the unit of observation and the

significance of the p-values is thus distorted. Where the total number of disagreements was less than 10, the exact p-value was calculated by assuming the number of disagreements in each direction followed a binomial distribution with probability 0.5.

Results

Series 04/06

Tables which compare relevant items of the surgeon's assessment and the patient questionnaire are shown in Table 11-1 and Appendix 9 (tables A9(i) to (vii)). Summaries of the agreement frequencies are in Table 11-2. No clear biases towards optimism or pessimism by the surgeon are obvious apart from for pain in activity for which the surgeon appeared optimistic. Table 11-1 shows that of 37 knees for whom the surgeon proclaimed "no pain", 16 patients marked a value of 1 or more on the visual analogue line.

Series 05

The results are shown in Tables 11-3, 11-4 and Appendix 9 (Tables A9(viii) to (xiv)) while agreement frequencies are shown in Table 11-5. A significant tendency towards optimism by the research registrar was seen for every item. The most compatible item concerned walking distance (Table 11-3). For example, of 10 patients who said they walked indoors only, only 4 were recorded as housebound by the registrar; the others were recorded as able to walk distances ranging from up to 400 metres to "unlimited". The converse did not hold: of the 7 patients recorded as housebound by the surgeon, 4 patients said they walked indoors only and three for up to 10 minutes. Concerning usage of a walking stick (Table 11-4), 12 patients agreed with the surgeon that they never used one. However 18 patients said they used one sometimes or often while the surgeon said they did not use one. Of those identified by the surgeon as using a stick, all patients agreed.

In summary either the patients in series 05 were exceedingly pessimistic when answering

a postal questionnaire, or the research registrar was consistently overrating the patients' experience.

Series 07

Table 11-6 and Appendix 9 (Tables A9(xv) to (xx)) shows all cross tabulations of surgeons' and patients' assessments while Table 11-7 shows agreement frequencies. Once again there was a consistent tendency for the surgeon to be more optimistic in the assessment of stair climbing, walking and pain during activity. Conversely however the surgeons were relatively pessimistic about whether the operation was worthwhile; of 46 patients who gave a score of 6 on the visual analogue score, only 24 were rated enthusiastic by the surgeon.

Overall surgeons appear to have overestimated patients' functional abilities and pain relief. The surgeon optimism does not appear as marked as for Series 05.

Series 08

There were few items where direct comparison could be made since this surgeon had concocted items of assessment which are slightly unconventional. Cross tabulations are shown in Table 11-8, 11-9 and Appendix 9 (Tables A9(xxi) to (xxiv)).

The agreement frequencies of Table 11-10 again suggest that patients were less optimistic than the surgeon. However the differences, although consistently in one direction, were not of great magnitude (see Appendix 9). The clearest differences were for pain assessment; of 28 knees considered by the surgeon to be giving no pain, some degree of pain was reported for 17 at rest (Table 11-8) and 13 knees in activity (Table 11-9).

Overall surgeon S8 was inclined to rate the outcome higher than were his patients in postal questionnaires. The effect was less marked for his own "doctor's assessment", where his use of radiological data induced an apparent cooling of enthusiasm.

Discussion

The results above show generally that clinical assessment leads to an apparently higher rating of outcome than does a patient's response to a postal questionnaire. This parallels the findings of Brewster and Newman (1991), as well as those of Black et al (1991) and Carlson and Pelletieri (1989). The present study has methodological problems, but these should not have spuriously created the findings.

Some assessments were made by surgeons up to 18 months prior to the questionnaire (Series 07,08). Given that intra subject variability was substantial for some items of the questionnaire (see Chapter 9), lack of agreement is only to be expected. However there is no reason to suppose that the direction of the disagreement is explained by the time gap. The analysis of Chapter 9 suggested that disagreements occurring over the space of one year were in no particular direction, and indeed such a deterioration over a relatively short average time gap is implausible. In addition the series where surgeon's and patients' assessments were made closest together in time (Series 05) showed most clearly the effect of "surgeon's optimism".

The definitions given of agreement frequencies are arbitrary. Tables of raw data sometimes suggest that equally reasonable alternative definitions would have lessened the evidence for "surgical optimism". However the items where direct comparison was easy uphold the general picture.

Brewster and Newman concluded that because patients did not agree with therapists that therefore the patients were wrong. Had the patients been answering haphazardly, the systematic tendency for patients to rate outcome less highly is unlikely to have been demonstrated. The above results confirm the unease that some orthopaedic surgeons have felt (see Bellamy and Campbell 1989). Both Black et al and Carlson and Pelletieri have claimed reasonably good agreement between surgeons and patients. Both their studies confirm however, that when disagreements occurred, patients' assessments tended to provide a more pessimistic view. Carlson and Pelletieri showed agreement was substantially better for assessments carried out on the day of discharge (when the patients

were still in hospital) compared with those done 8-24 months post discharge (when the patients filled in a postal questionnaire). Black et al showed substantially less agreement for symptoms than for complications. Brewster and Newman claimed there was poor agreement but because their items were based on a score rather than on a dichotomous response (Black et al) or a grade of 1 to 5 (Carlson and Pelletieri), it is difficult to compare easily the extent of disagreement in the three studies. There is however a consistent pattern in the direction of disagreement.

It would equally be unfair to orthopaedic surgeons to suggest they were somehow wrong in their assessments. Rather, it is likely that they were using different criteria by which to judge. Patients filling in questionnaires anonymously within their own homes may have been able to be more realistic about every day experience while when faced with a member of the surgical team they may have been more inclined to follow "normative effects" (Fitzpatrick 1990). A study of 88 hip replacement patients (Burton et al 1979) found that 86% declared the operation successful but only 55% said their expectations had been fulfilled. Surgeons may be more interested in technical success of the operation and be concerned to assess the **maximum** functional capabilities of which the patient is capable. It may also be argued that patients' apparent pessimism was linked to their own unrealistic expectations (Streetly and Morris 1992). If patient judged outcome results in undue pessimism, this bias may in fact be preferable for the purpose of medical audit (Black et al 1991).

There was some evidence that the difference between surgeon's and patient's assessment was greater for some series than others. If it is assumed that differences were entirely the fault of the patient, this would have implications for the comparison of series as made in Chapter 10. For example in series 05, which ranked low in its median knee score, a particular divergence was seen. It could be argued that the results in terms of knee score for series 05 paint an unfair picture. However in series 04 and 06, which rank low in terms of median knee score, relatively good agreement between surgeon and patient was seen. Conversely in series 07, where a high median knee score was obtained, there was evidence that patients were being pessimistic relative to the surgeon. Thus the high rank of series 07 is further vindicated.

Nonetheless it is implausible that the differences seen in this study and in the three studies in the literature, were entirely the fault of the patient. The magnitude of disagreement clearly varied between series, and this is more likely to be explained by the different assessment criteria of the different surgeons rather than systematic variation in responses to a pre-printed questionnaire by groups of patients treated by different surgeons. The futility of comparing results from different case series reported in the literature is underlined.

Methods of assessing the outcome of knee replacements by surgeons within the outpatient clinic were designed in the 1970s (for example Insall et al 1976, Aichroth et al 1978). These have changed very little up to the present. They were designed for an era when knee replacements were still being pioneered. Given the very high success rates quoted in present literature even at 12 years or more post operation, the time may have come to reappraise the whole philosophy of outcome assessment. The present study has confirmed that not every patient agrees with a surgeon's pronouncement of an excellent result. Furthermore, patient responses demonstrate greater variation between subjects than do surgeon assessments. This is important for any outcome measure designed to demonstrate differences in success of different prostheses.

Conclusion

Patients' answers to anonymous questionnaires give less rosy ratings of outcome of knee replacement than do surgeons' clinical assessments made in outpatient clinics. Whilst long term clinical follow up is still necessary to ensure technical quality of results, postal questionnaires may provide the basis for a form of audit more sensitive to patients' perceived needs, and more capable of showing differences in outcome between groups of patients treated by different surgeons with different prostheses.

Table 11-1. Comparison of surgeons' and patients' assessments of pain in activity (knees) (Series 04/06)

Patient

Whilst moving around, how much pain have you had from this knee in the past week?

Surgeon

Severity of pain	0	1	2	3	4	5	6
None	21*	8	3	1	4	0	0
Mild	0	0*	1*	2	1	0	0
Significant	3	0	0	0*	3*	0	0
Severe	0	0	0	0	0	0*	0*

* shows cells of table representing agreement between surgeon and patient.

Table 11-2. Agreement frequencies between surgeons' and patients' opinions (Series 04/06)

Item	Agree	Surgeon pessimistic	Surgeon optimistic	p-value *
Rising from chair	18	4	10	0.14
Going upstairs	16	12	4	0.06
Going downstairs	18	8	6	0.40
Walking distance	16	5	11	0.17
Use of walking cane	11	11	3	0.07
Pain in activity	25	3	19	<0.001
Worthwhile operation	36	7	2	0.13

* using McNemar's test for asymmetry of disagreement

Table 11-3. Comparison of surgeons' and patients' assessments of walking distance (Series 05)

	Patient					
Surgeon	Not at all	Indoors only	Up to 10 mins	10-30 mins	30-60 mins	Over 60 mins
Unable	0*	0	1	0	0	0
Housebound	0	4*	3	0	0	0
<400 m	0	2	7*	3*	0	0
400-800m	0	2	6	2	0*	1
>800m	0	1	0	5	3	2*
Unlimited	0	1	1	4	5	6*

* shows cells of table representing agreement between surgeon and patient.

Table 11-4. Comparison of surgeons' and patients' assessments of use of walking sticks (Series 05)

	Patient		
Surgeon	Never	Sometimes	Often
No stick	12*	6	12
One stick	0	4*	16*
Two sticks	0	1*	0*
Crutches/ frame	0	0*	2*

* shows cells of table representing agreement between surgeon and patient.

Table 11-5. Agreement frequencies between surgeons' and patients opinions (Series 05)

Item	Agree	Surgeon pessimistic	Surgeon optimistic	p-value *
Rising from chair	35	4	21	<0.001
Going upstairs	37	4	17	0.009
Descending stairs	32	1	26	<0.001
Walking distance	22	5	32	<0.001
Walking sticks	35	0	18	<0.001
Pain at rest	51	4	29	<0.001
Pain in activity	47	11	26	0.02

* using McNemar's test for asymmetry of disagreement

Table 11-6. Comparing surgeons' and patients' assessment of whether the operation was worthwhile (Series 07)

Patient

How worthwhile do you feel that this knee replacement has been?

Surgeon

	0	1	2	3	4	5	6
Disappointed	0*	0*	0	0	0	0	3
Non-committal	0	0	0*	0*	1	0	1
Satisfied	0	0	0	2	0*	4*	18
Enthusiastic	0	0	0	1	2	6	24*

* shows cells of table representing agreement between surgeon and patient.

Table 11-7. Agreement frequencies between surgeons' and patients' opinions (Series 07)

Item	Agree	Surgeon pessimistic	Surgeon optimistic	p-value *
Rising from chair	46	19	11	0.20
Going upstairs	45	5	27	<0.001
Walking distance	31	16	39	0.003
Pain at rest	37	6	36	<0.001
Pain in activity	32	6	37	<0.001
Worthwhile operation	28	23	11	0.06

* using McNemar's test for asymmetry of disagreement

Table 11-8. Comparison of surgeons' and patients' assessments of resting pain (Series 08)

Patient

Whilst resting, how much pain have you had from this knee in the past week?

Surgeon

	0	1	2	3	4	5	6
None	11*	11	3	3	0	0	0
Slight	2	3*	1*	2	0	0	1
Moderate remittant	0	0	0	0*	0	0	0
Moderate constant	0	0	0	0	0*	0	0
Severe remittant	0	0	0	0	0	0*	0
Severe constant	0	0	0	0	0	0	0*

* shows cells of table representing agreement between surgeon and patient.

Table 11-9. Comparison of surgeons' and patients' assessments of pain in activity (Series 08)

Patient

Whilst moving around, how much pain have you had from this knee in the past week?

Surgeon

	0	1	2	3	4	5	6
None	15*	7	3	3	0	0	0
Slight	4	1*	2*	1	0	1	0
Moderate remittant	0	0	0	0*	0	0	0
Moderate constant	0	0	0	0	0*	0	0
Severe remittant	0	0	0	0	0	0*	0
Severe constant	0	0	0	0	0	0	0*

* shows cells of table representing agreement between surgeon and patient.

Table 11-10. Agreement frequencies between patient' and surgeon's opinions (Series 08).

Item	Agree	Surgeon pessimistic	Surgeon optimistic	p-value *
Worthwhile operation	15	2	20	<0.001
Pain at rest	18	4	15	0.02
Pain in activity	24	3	11	0.06
"Doctor's assessment"	27	3	11	0.06

* using McNemar's test for asymmetry of disagreement

CHAPTER 12. APPLICATION OF FINDINGS TO FUTURE EVALUATIVE RESEARCH ON OUTCOMES OF KNEE REPLACEMENT

Introduction

This chapter will review different methods of measuring outcome of knee replacements, and consider the use of the postal questionnaire. It will also look at different possible types of study design in evaluating the efficacy of different prostheses. In particular the possibility of mounting randomised controlled trials will be considered, and various issues in randomisation will be explored.

Recommendations for types of statistical analysis will depend on the design of studies and the outcome measures chosen. I have contributed elsewhere to discussing the use of survival analysis (Lettin et al 1991, Morris 1992a, Carr et al 1993a). This has included a recommendation to calculate confidence intervals for survival probabilities based on the formula of Peto et al (1977), to consider carefully the definition of endpoints and to consider the potential bias of those lost to follow up. Carr et al produced a checklist for researchers organising a trial of a new form of joint replacement.

I have also advised how best to handle observations on patients with both knees replaced (Morris 1993b). This has been considered in Chapter 5, where only the first knee replaced was used for comparative analysis of patients treated with two different prosthesis subtypes. In Chapter 10, where a quantitative knee score was used for analysis, the average of two knees in bilaterally replaced patients was taken. A technique to give greater weight to data from such patients was used but this is probably unnecessary, given the relatively small between knee variance within patients.

The remainder of this chapter will consider the more fundamental questions of outcome measurement and study design in future research.

OUTCOME

Are surgeon based measures of outcome adequate?

Over the two decades or more since knee replacement began to be carried out by increasing numbers of surgeons, three types of outcome have been used in publishing estimating success. These have been rates of reoperation (survival analysis), clinical assessment and radiographic assessment. The suggested changes in defining the last two, made by Insall (1989) and Ewald (1989), have not altered the basic philosophy. Clinical assessment has been slanted towards the detection of untoward events (such as loosening or infection of the prosthesis). In recent years, reported survival rates (taken at face value) have been so high that there would no longer seem scope for improvement.

Chapter 5 of this study has suggested that alternative plausible definitions of failure lead to differing survival rates. One of the chief difficulties, recognised for many years by some authors (eg Tew and Waugh 1982) is that a knee replacement which is painful but not revised because of other medical reasons, is not counted a failure if survival is taken as meaning lack of revision surgery. This study suggests a three fold increase in failure rates may occur with alternative outcome definitions.

Concerning clinical assessment, Chapter 11 has demonstrated that surgeons are consistently liable to overestimate patients' functional capabilities and pain relief. Although pessimism by the patients is also a plausible explanation, it would appear that surgeons find it difficult to understand patients' subjective feelings about their functional capabilities and pain experienced in their replaced knee joint.

Radiographic assessment has not been investigated in this thesis, but other workers have suggested that the association between radiographic and clinical outcome is disappointingly weak (Tew and Waugh 1985). The roentgen-stereophotogrammetric analysis (reported by Grewal et al 1992) is as yet unproven. At the moment no reliable measurement may be taken to predict long term survival.

What perspective does the patient questionnaire provide?

Since the analysis of Chapter 11 suggested that patients may not view outcome as optimistically as the surgeons, the estimate of success derived from patients' questionnaires is of interest. In a satisfaction score which could range from 0 to 42, no less than 20% of patients returned the perfect score of 42, while the median score was 35. This suggests that this sample of patients was generally satisfied with its replaced knees.

However satisfaction varied according to a number of factors. Some series had lower satisfaction ratings, in particular an old series (04/06) which dated from the 1970s. This series used the hinged prosthesis, which has been seen as of very little use in the 1990s (Noble and Hilton 1991). The other three of the old series had very high ratings, among them series 08. However another surgeon who used the same prosthesis obtained poor results (Series 05). The results of Series 07 were nearly matched by another more recent series which used the same prosthesis. The sum of evidence was that pioneer surgeons were inclined to outperform others who used their prostheses.

As may have been expected, other factors related to satisfaction expressed by the patients. RA patients were more satisfied than OA patients, perhaps because their more restricted lifestyle makes fewer demands on their knee replacement. Those often using aids were less satisfied as were patients residing in old people's homes at the time of filling in the questionnaire. Satisfaction was less among patients who had undergone their knee replacement a long time ago. However these differences did not explain away the difference between the surgeon series. The odds of a "poor" result (satisfaction score 35 or less) among patients in the worst series was 5 times that for those in the best series.

Issues in using a patient questionnaire

(a) Magnitude of differences in patient satisfaction

Table 10-1 (Chapter 10) showed means and standard deviations of the knee score for each

Series. Good series had means of around 33 while poor series averaged around 29.5. The standard deviation for every series was around 10. Thus to demonstrate statistically significant differences at the 5% level between good and poor series with 90% power, future studies would require **171** patients per group. However questions of future interest in comparing groups of patients undergoing alternative forms of knee replacement would probably involve comparing a potentially "good" series (mean knee score of 33) with at worst an average series (mean knee score of 31,say). In this situation around **525** patients per group would be required. Although this sounds a hard task, it is small compared with the numbers required to demonstrate differences in survival rates at ten years. For example the comparison made by Scuderi et al (1989) of Total Condylar Prostheses with or without metal backed tibial components showed 7 year survival rates of 97.34% and 98.75% respectively. The numbers of patients required to demonstrate this magnitude of difference would be in excess of 2000 per group.

(b) Response rate

The response rate in this cross sectional study was only 60% of all those who had undergone knee replacement. Almost half of the non response was accounted for by dead patients. Others were not known at the address given. However 13% of the patients did not reply for unknown reasons. To improve response rates, dedicated full time coordinators would be required to chase up non responders. Nonetheless, to have gained some reply from 960 out of 1166 patients who would have received the questionnaire as far as we may assume, is encouraging. The fact that 70% of patients wrote comments at the end of the questionnaire suggests that they were interested and motivated; indeed some patients even offered spontaneously to be interviewed about their knee replacement!

Any prospective study would involve regular contact with the patient cohort, so changes of address would be known about. Of those patients still alive, response rates in excess of 90% would be feasible.

(c) Time after operation

It was shown in Chapter 10 that differences between the series were visible after follow up periods of 0-3 years. If this result may be generalised, it suggests a relatively short follow up would be necessary in prospective studies. Of course this would minimise loss of patients due to death or changes of address.

(d) Reliability

Chapter 9 noted some lack of agreement in replies to a questionnaire administered on two occasions one month apart. This was especially the case for a number of functional items. Those related to getting out of a bed and rising from a chair and phrasing of these questions may need to be tightened up for future studies. The same was true for use of aids, in particular defining the amount of usage. Although items on pain per se were not very reliable, the amount of agreement in total knee score was encouraging.

The questionnaire was photocopied on two sides and some patients seemed to miss out the "back" pages. If the total length of the questionnaire can be reduced, single sided pages may be worthwhile.

(e) Validity

There are no external criteria for validity of replies when subjective measures are involved. It has been noted that patients were less optimistic in their assessments than surgeons but a careful investigation of the cross tabulations in Appendices 11-1 to 11-4 suggests there was at least some correlation between the two. The poor results found for Series 04/06 are in accord with the apparent consensus on the relatively poor results that the hinged knee replacement gives (Walker 1989, Noble and Hilton 1991). The relationship of various factors to knee satisfaction score are not unexpected.

(f) Standard for all patients

One of the criticisms made in Chapter 3 of past research on knee replacements is that definition of outcome must inevitably vary between surgeons. In this study every series of patients has been confronted with the same preprinted questionnaire. Bias could only be created if the patients of different surgeons varied in their reactions to the questions. This possibility was diminished by assuring patients that no doctor involved in caring for them would see the replies they made. Although this does not eliminate the possibility that patients have assigned higher scores to more charming surgeons, it is an improvement on the situation where the surgeon himself assesses the patient in an outpatient clinic.

(g) Dimensions

Most published scales which assess quality of life are organised into different dimensions, whereby questions relating to one particular aspect of life are grouped together. The knee score used in Chapter 10 could be criticised because it aggregates different dimensions in a simple way. Bellamy and Campbell (1989) disagreed with aggregation. However in any clinical trial, there is a need to identify a small number of outcome variables (ideally one), to avoid repeated use of significance tests. The present study showed that values from different items tended to follow similar patterns within patients.

The WOMAC questionnaire

The WOMAC instrument developed in Canada and used by Bellamy et al (1988) for evaluating hip and knee replacements seems little known among the British orthopaedic community. However it seems very promising and once certain items have been reworded for a British population, it may have some attraction for future usage. It has been tested more fully than the questionnaire used in this thesis.

STUDY DESIGN FOR COMPARATIVE RESEARCH

Chapter 3 drew attention to the large degree of usage of case series in evaluating knee replacements, and commented on the great difficulties in making reliable comparisons. These may have been appropriate in an era when large improvements were more likely in successive case series. Now the improvements in success rates are likely to be only modest and the size of biases created by comparing case series will cause a serious problem. Consideration must therefore be given to the possibility of mounting randomised trials.

Randomised trials and knee replacements

It has been established and recognised over the past four or five decades that the best method for evaluating new therapies in medicine is the randomised controlled trial (RCT). The purpose is to produce two or more groups of patients in whom the only differing characteristic is the treatment group to which they were assigned. Thus any difference in outcome between the groups should be imputed to the treatment itself, unless imbalance of the groups of patients with regard to relevant characteristics has occurred by chance.

A trial should be undertaken when there is a genuine doubt as to which of two treatments is preferable for the defined group of patients. One of the thorniest issues concerns the notion of "informed consent". The judgement that an individual patient is eligible for the trial should, according to the Helsinki declaration, include explaining to the patient the nature of the trial and the treatment they are undergoing, and obtaining consent at least in verbal, but preferably in written form.

History of RCTs

The first randomised trial in the modern sense concerned the MRC's evaluation of streptomycin for pulmonary tuberculosis in the late 1940s. Since then, randomised trials have formed a growing industry and increasing numbers of trials are published in the

medical literature. Their most favoured application has been in the realm of drug evaluation, and has been thoroughly expounded by Pocock (1983).

Surgeons have indeed also entered into this whole paradigm (though more cautiously). The first serious randomised trial (Goligher et al 1964) compared three alternative forms of surgery for patients with duodenal ulcers and the trial is now thought as having established the efficacy of vagotomy for such patients.

According to Pollock (1989), there were no RCTs published in the British Journal of Surgery in 1965, but in 1985 there were 28. Such trials may involve the testing of alternative ancillary prophylactic measures used in conjunction with the surgery itself, for example the use of anticoagulants to prevent deep vein thrombosis after joint replacement. However surgical trials may involve testing a surgical procedure against a pharmacological alternative, for example the trial comparing extracranial - intercranial bypass surgery for preventing strokes (Sundt 1987).

More relevantly for the current thesis, trials such as the vagotomy study have been carried out which involve alternative surgical procedures. However, as was reported in Chapter 3 of this thesis, not a single report of a randomised controlled trial could be identified concerning the world literature on knee prostheses listed in Index Medicus for 1987 to 1989. In fairness however this ethos may be changing slowly. Nilsson et al (1991) carried out a small randomised study of cemented and uncemented knee replacements, making RSA measurements after a short postoperative interval. The study involved only 43 patients, scarcely enough to make a proper comparison, and points to the formidable barriers to carrying out proper randomised trials. This is presumably because knee surgeons have decided either consciously or subconsciously that such research methodology is unnecessary, not feasible or both. There are however a number of partially unresolved issues which may be worth considering.

1. Knee surgery versus conservative management

There are a whole spectrum of conservative treatments available to clinicians other than

surgery, and this may include use of a walking stick, advice to lose weight, or steroidal and non steroidal anti inflammatory drugs prescribed by rheumatologists. Many surgeons who carry out knee replacement, however, do so when these conservative remedies appear to have failed (Noble and Hilton 1991). At this stage it may be argued that to randomise patients to continuing conservative management as compared with surgery is unethical. However waiting lists for elective surgery have been a hot political issue in recent years (Yates 1991) and orthopaedic surgeons have frequently been denied the theatre time necessary to meet all the perceived need. Knee replacements are bound to be a prime example. If this is the case, then some form of rationing is inevitable. Evidence of underprovision of services was found by Wallace et al (1992) who surveyed a general practice near to their orthopaedic department in Nottingham. In the practice list they found patients on their computerised diagnostic register with diagnosis of osteoarthritis of the knees. Of 52 patients, only two had undergone knee replacement and of the remaining 50, 18 were considered eligible. It is thought that GPs are still sceptical as to the success of knee replacements, but if such scepticism is dispersed, demand for knee replacement could burgeon. When the MRC carried out its trial of streptomycin in the 1940s, one of the ethical arguments in favour of the trial was that the drug was then in short supply. Therefore since not all patients with pulmonary TB could receive streptomycin anyway, randomisation was a fair way of deciding.

There may therefore be a case for randomly allocating patients who need knee replacement to being placed on a year's waiting list or being treated immediately. At the end of the year some standard assessment could be made of knee function and knee pain for the patient whichever treatment "arm" (s)he had been assigned to. The option of randomising patients to immediate or late treatment will be explored further below.

Of course such a scheme can only provide a quick short term glimpse of the effect of knee replacement. Furthermore, the necessity for waiting lists may suddenly dissolve for a period within a particular hospital, and this would alter the ethics of the situation dramatically. The situation with regard to hospital waiting lists prevalent at the outset of such a study would have to remain constant throughout the period of accrual of the necessary number of patients to the study.

This type of study may be of use to departments of Public Health in their new role of advising the purchasers of local health services. It has the attraction of containing immediate relevance. However it can scarcely provide a full scientific answer to the question concerning efficacy of knee replacement.

2. Knee replacement versus other surgical procedures

If it is accepted that knee surgery is needed for a defined group of arthritic patients, the best surgical procedure is to be decided upon. As was mentioned in Chapter 1, high tibial osteotomy was considered a viable alternative to correct osteoarthritis manifesting on the medial side of the knee. Also arthrodesis has been used to treat severe cases of arthritis. The current view of orthopaedic surgeons is that each of these options is very outdated. It is admitted however that osteotomy may be useful for younger patients in order to preserve bone stock. Meanwhile arthrodesis is seen as a final option, after knee replacement itself has failed. Thus although there exist small subgroups of patients in whom genuine questions about alternative surgical options exist, it is highly unlikely that such a trial would arouse sufficient interest or commitment from surgeons to accrue required numbers of patients.

3. Comparison of alternative types of knee replacement

Since knee replacement has become a widely used procedure in Britain and other western countries, there should exist sufficient numbers of patients to carry out a trial. Goddard and Coleman (1992) reported that 32 different prostheses are in current use. A lack of consensus seems to exist concerning which is the best, and Bulstrode et al (1993) suggested some joint replacement patients were victims of fashion. Meanwhile manufacturers produce new prostheses at frequent intervals, and these prostheses may be introduced into immediate clinical practice. Unlike pharmacological agents, these prostheses have not required the approval of a government monitoring body such as the Committee of Safety on Medicines. Challah and Mays (1986) complained of double standards with reference to the introduction of new medical technology. Recently a government report has recommended a Committee on Safety and Efficacy of Procedures

(Advisory Council on Science and Technology 1993). Meanwhile Gross (1993) has argued for a four phased testing procedure of new devices, analogous to a standard of drug evaluation. Goodfellow (1993) has supported the need for randomised trials.

Varieties of knee replacements have evolved over the past twenty years, but the main varieties which are still in contention are the minimally constrained total condylar type and the unicompartmental variety. Secondary to this is a debate about the need to retain the cruciate ligaments to maintain balance of the joints, or whether to design a prosthesis which carries out this function.

So could trials be organised so that patients requiring knee replacement would be randomly allocated to a total condylar versus a unicompartmental prosthesis ? Clearly even among the general types of prosthesis identified above there exists variety. This need involve no difficulty since the surgeon involved could choose the prosthesis he then feels most comfortable with. However there are still a number of difficult barriers to be surmounted and choices to be made.

a) Indications for surgery

Communications from surgeons participating in the study described in this thesis have made it clear that they hold different selection criteria. For example the Oxford knee (meniscal) is nowadays applied only to patients in whom the cruciate ligaments are intact (Goodfellow et al 1988). In practice this means patients in the earlier stages of osteoarthritis, attacking only one side of the knee joint. One surgeon professed disapproval of unicompartmental prostheses, "I prefer to wait a couple of years and then go for a full total". This suggests that early treatment with a total semi constrained knee replacement is disliked by surgeons. Conversely, unicompartmental meniscal prostheses should be inserted at an early stage to gain the best results (Thornhill and Scott 1989). Carr et al (1993b) suggested early surgery might prevent arthritis spreading across the knee joint. The possibility therefore arises again that a randomised trial comparing early and late surgery could be feasible. This may be more acceptable than forcing on surgeons one prosthesis rather than another for patients in immediate need of surgery. Instead the

dilemma of operating now rather than later, in patients for whom a genuine choice must be made, may be broached by randomisation. Since surgeons are likely to find an ever increasing pool of patients asking for knee replacement, and resources will eventually be saturated, such a question has a wide application in health service provision.

Such a randomised study would therefore need to identify patients at an early stage of arthritis of the knee joint. The allocation would be either to receiving an immediate unicompartmental prosthesis, or to waiting a given period of time and being reassessed for the need of knee replacement of the condylar variety. Only surgeons comfortable with either option could be recruited.

Some of the patients allocated to the second arm of the study may never in fact undergo knee replacement with a condylar prosthesis. The trial question would therefore be "Is unicompartmental knee replacement for early arthritis of the knee better than conservative management?". Any assessment of outcome would need to be related to patient status before entry to the trial, not before surgery. This would be applicable to patients allocated to the second arm, even if they never receive a knee replacement. The trial would fall into the so-called "pragmatic" category rather than being "explanatory", to use the terms coined by Schwartz et al. (1980). This would appear to be a question of genuine practical interest to orthopaedic surgeons.

This proposal would cut rheumatoid arthritis patients out of the reckoning, since unicompartmental knee replacement is inappropriate because of the nature of their disease. However the concept of testing immediate total knee replacement against delaying for two years may still be plausible for patients within the early stages of RA.

b) Treatment allocation

It has been explained above that surgeons have historically shown markedly less enthusiasm for randomised controlled trials than have physicians. The chief reasons for this appear to be:-

(i) Interference with clinical freedom. It has been argued that surgeons rarely feel neutrality when weighing alternative treatment options for individual patients. Random allocation would thus compromise the surgeon's integrity.

(ii) Unease in obtaining informed consent from patients. Surgical procedures have more dramatic consequences than do consumption of tablets (Love 1975). Thus explaining to patients that their treatment will be decided on the basis of chance rather than on surgical expertise is liable to compromise the doctor - patient relationship (Dudley 1985, 1991). Besides this, present culture tends towards respecting patient autonomy, and it has been suggested that patients are less likely to enter the trial when full information is given to them (Lancet 1991b).

(iii) Differing levels of surgical skill. If surgeons are asked to carry out alternative procedures on a given group of patients, it is required that they perform each with the same level of skill. Unfortunately this is rarely the case in practice. Most new surgical procedures involve a "learning curve", and knee replacement is no exception. Traditionally it has been carried out by consultants only, though this is less true now.

Conventional randomised trials in recent years have run into formidable obstacles. Considerable doubts were raised over the EC/IC Bypass study when it was realised that less than half the eligible patients had entered the study (Sundt 1987). In 1982 the Cancer Research Campaign set up a study of total mastectomy versus segmental for breast cancer, but abandoned it after 3 years when only 145 patients had been recruited (see Pollock 1989). A proposed randomised study of Extra Corporeal Shock Wave Lithotripsy versus open or percutaneous surgery for the removal of renal stones was refused after considerable opposition from urologists (Challah and Mays 1986). More recently, Neugebauer et al (1991) decided not to proceed with a randomised study of conventional versus laporoscopic cholecystectomy. They had felt it wrong to start with a randomised study while the new laporoscopic technique was still being learned by surgeons. The pilot study of the new technique however then gave results that were apparently too encouraging to contemplate continuing with a randomised study.

A number of authors have grappled with special design requirements for randomised trials in orthopaedics (Laupacis et al 1989, Rudicel and Esdaile 1985, Chang et al 1990). It would appear nonetheless that considerable willpower of orthopaedic surgeons would be required to make the study of unicompartmental knee replacement proposed above succeed. The only hope of assuaging the ethical objections of both surgeons and patients must lie in the existence of a genuine need for rationing the number of patients on whom knee replacement can be carried out.

Surgical skill

Rudicel and Esdaile (1985) fully thrashed out the concept of the "randomise to surgeon" design. Instead of individual surgeons being required to alternate between different procedures, it was suggested that each surgeon stick to the procedure with which (s)he feels most comfortable. Random allocation would then be to the surgeon responsible for carrying out the assigned procedure. The idea was rebutted in a Lancet editorial (1986a) which called for a return to conventional randomisation. Once again the issue was surgical skill. Paradoxically, Rudicel and Esdaile seemed to magnify the very problem they sought to solve. Any difference in outcome may, with their design, be imputed to the surgeons rather than the treatment per se. In addition the design seems to resemble "randomisation by cluster" where homogeneous blocks of patients are assigned to be treated by a given surgeon in a given way. The relative efficiency of this design, following Cornfield (1978), would decrease with the magnitude of the surgeon effect, but would increase with the ratio of surgeons to patients participating. To avoid changing the required sample size, given the surgeon effect shown in Chapter 10, no more than 130 patients per surgeon should be included (see Appendix 10). This implies at least four surgeons would be required per treatment group.

Which prostheses?

If the trial used only one prosthesis of each variety, the applicability of the findings would be clear but restricted. If however surgeons used the choice of his own unicompartmental prosthesis, the application would be less clear for those interested in

evaluating efficacy of alternative prosthesis models. In the interests of surgeons' compliance in such a study, the latter option would be preferable. In addition the results of a trial using the latter option would be of interest not only to surgeons but also to purchasers of health care and in general to all parties concerned with health service provision for this patient population.

Which surgeons?

Chapter 10 showed the definite existence of a surgeon effect in outcome for a given prosthesis. In particular the surgeon who invented the prosthesis was liable to obtain better results than other surgeons using the same prosthesis. This suggests that prosthesis pioneers should not be included in the study. The results they obtain will produce an over optimistic picture. Conversely however, since knee replacement presumably involves a learning curve, inexperienced surgeons should not be involved either.

A mail survey to orthopaedic surgeons explaining the purpose of such a study and asking for their participation is likely to be necessary. This would be coupled with adverts at orthopaedic meetings and individual phonecalls by the study coordinator. A proactive approach would be required to attract sufficient surgeons. No information is nationally available on numbers of patients who are currently candidates for unicompartmental replacement, nor on numbers of surgeons happy to carry out such a procedure.

It has been argued above that over 500 patients per treatment would be required for such a trial, involving at least four surgeons per treatment. In any case the only feasible way to recruit the patients within a reasonable period of time would be to use several surgeons.

What outcome variable ?

A patient questionnaire seems to be a sensitive way of picking up differences between treatment groups. However surgical assessment at regular intervals is also important. This would fulfil the customary obligation to the patients' welfare as well as monitoring

survival rates of the implanted prostheses.

The further question concerns the time point of any questionnaire. Information on the "early surgery" group and the "delay" group would be helpful during the first two years. However five years after random allocation would also be a good time, and this could be taken as the main outcome time point. However in order to retain contact with patients, brief communications at yearly intervals would also be helpful.

Conclusion

Conventional forms of clinical assessment, whilst of importance in auditing technical aspects of the outcome of knee replacement, are now too insensitive to be of benefit in comparative research on alternative prostheses. Confidential patient questionnaires appear to be more suitable for uncovering remaining problems in this era of high surgical success rates. Given the relatively modest improvements in outcome to be expected with newer forms of prosthesis, comparative studies which use randomisation will be more credible than studies involving case series, which are subject to many biases.

Randomised controlled trials for alternative surgical procedures have not so far met with much enthusiasm from the surgical community. Situations where rationing of treatment is inevitable may override the ethical barriers to randomisation. This may be the case for those patients referred to surgeons with early arthritis attacking one side of the knee joint. A successful trial is likely to require the cooperation of many surgeons in order to realise 1000 patients. A postal questionnaire is recommended as a sensitive measure of outcome.

REFERENCES

Acta Scandinavica Orthopaedica (1988) Multicenter investigations. The role of orthopaedic journals (editorial). Acta Orthop Scand 59: 247-8.

Advisory Council on Science and Technology (1993) A report on medical research and health. London : HMSO

Aglietti P, Buzzi R (1988) Posteriorly stabilised total condylar knee replacement. Three to eight years' follow up of 85 knees. J Bone Joint Surg [Br] 70-B: 211-6.

Agresti A (1990) Categorical data analysis. New York: Wiley

Aichroth P, Freeman MAR, Smillie IS, Souter WA (1978) A knee function assessment chart. J Bone Joint Surg [Br] 60-B : 308-9.

Albrektsson BEJ, Ryd L, Carlsson LV, Freeman MAR, Herberts P, Regner L, Selvik G. (1990) The effect of a stem on the tibial component of knee arthroplasty. A roentgen stereophotogrammetric study of uncemented tibial components in the Freeman-Samuelson knee arthroplasty. J Bone Joint Surgery [Br] 72-B: 252-8.

Altman DG (1991a) Practical statistics for medical research. London : Chapman & Hall

Altman DG (1991b) Statistics in medical journals: developments in the 1980s. Statistics in Medicine 10: 1897-1913

Armitage P, Berry G (1987) Statistical methods in medical research. 2nd edition. Oxford : Blackwell

Apley AG (1990) An assessment of assessment. J Bone Joint Surg [Br] 72-B: 957-8.

Attenborough CG (1978) The Attenborough total knee replacement J Bone Joint Surg

[Br]; 60-B: 320-6.

Barrett DS, Biswas SP, MacKenny RP (1990) The Oxford knee replacement. J Bone Joint Surg; 72-B: 775-8.

Bebbington AC (1977) Scaling indices of disablement. Br J Preven Soc Med 31: 122-6.

Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt L (1988) Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes following total hip or knee arthroplasty in osteoarthritis. J Orthop Rheumatol 1: 95-108.

Bellamy N, Campbell J. (1989) Hip and knee rating scales for total joint arthroplasty: a critical but constructive review. J Orthop Rheumatol 3; 3-21 (Part I), 63-76 (Part II).

Bergner M, Bobbit RA, Carter WB, Gilson BS (1981) The Sickness Impact Profile: development and final testing of a health status measure. Medical Care 19: 787-805.

Bishop YMM, Fienberg SE, Holland PW. (1975) Analysis of discrete multivariate data. London : MIT Press

Black N, Petticrew M, Ginzler M, Flood A, Smith J, Williams G, Davies J, Doll H, McPherson K (1991) Do Doctors and Patients Agree? Views of the outcome of transurethral resection of the prostate. Int J Technol Assess 7: 533-44.

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i: 307-10.

Brewster SF, Newman JH (1991) Can knee replacements be assessed by post? Health Trends 23: 113-4.

Bulstrode C (1987) Keeping up with orthopaedic epidemics. Br Med J; 295: 514

Bulstrode CJK, Murray DW, Carr AJ, Pynsent PB, Carter SR (1993) Designer hips. Br Med J 306: 732-3

Burton KE, Wright V, Richards J (1979) Patients' expectations in relation to the outcome of total hip replacement surgery. J Rheum Dis 38: 471-4.

Carlson H, Pelletieri L (1989) Doctors' versus patients' evaluation of results after neurosurgery. J Neurol Neurosurg Psychiat 52: 153-5.

Carr AJ, Morris RW, Pynsent PB, Murray DW (1993a) Survival analysis of joint replacements. J Bone Joint Surg [Br] 75-B: 178-82.

Carr A, Keyes G, Miller R, O' Connor J, Goodfellow J (1993b) Medial unicompartmental arthroplasty. A study of the Oxford meniscal knee. Clin Orthop (in press)

Challah S, Mays NB (1986) The randomised controlled trial in the evaluation of new technology : a case study. Br Med J 292: 877-8.

Chang RW, Falconer J, Stulberg SD, Arnold WJ, Dyer AR (1990) Prerandomization : an alternative to classic randomization. J Bone Joint Surg [Am] 72-A: 1451-5

Chinn S, Burney PGJ. (1987) On measuring repeatability of data from self-administered questionnaires. Int J Epidem 16: 121-7.

Cornfield J (1978) Randomisation by group : a formal analysis. Am J Epidem 108: 100-1.

Cox DR (1972) Regression models for life tables. J R Statist Soc (B) 34: 187-220.

Cox DR (1970) The analysis of binary data. London : Methuen

Cox DR and Oakes D (1984) Analysis of survival data. Chapman-Hall.

Dieppe P (1990) Osteoarthritis. A review. J Roy Coll Phys 24: 262-7.

Dixon WJ (1988) BMDP Statistical Software Manual. University of California Press

Dobbs HS (1980) Survivorship of total hip replacements . J Bone Joint Surg [Br] 62-B : 168-73.

Donner A, Donald A (1988) The statistical analyses of multiple binary measurements. J Clin Epidemiol 41: 899-905.

Dorey F, Amstutz HC (1989) The validity of survivorship analysis in total joint arthroplasty. J Bone Joint Surg [Am] 71-A: 544-7.

Drewett RF, Minns RJ, Sibly TF (1992) Measuring outcome of total knee replacement using quality of life indices. Ann Roy Coll Surg Eng 74: 286-90.

Dudley HAF (1985) Trials and tribulations for surgeons. Br J Surg 72: 255.

Dudley HAF (1991) The controlled clinical trial. Surgery 92: 2186 (add-on series)

Ewald FC (1989) The Knee Society total knee arthroplasty roentgenographic evaluation and scoring system. Clin Orth 248: 9-12

Felson DT (1990) Osteoarthritis. In : Hochberg MC ed. Epidemiology of rheumatic disease. Rheum Dis Clin N Amer 16: 499-512.

Fitzpatrick R, Newman S, Lamb R, Shipley M (1989) A comparison of measures of health status in rheumatoid arthritis. Br J Rheumatol 28: 201-6.

Fitzpatrick R (1990) Measurement of patient satisfaction. In: Measuring the outcomes of

medical care (eds. A Hopkins and D Costain). London: Royal College of Physicians.

Fox AJ, Goldblatt PO (1982) Longitudinal Study: socio-demographic mortality differentials: a first report on mortality in 1971-1975 according to 1971 Census characteristics, based on data collected in the OPCS Longitudinal Study. London : HMSO

Fleiss JL (1981) The analysis of cross classified data. New York : Wiley

Freeman MAR, Railton GT (1987) The current status of total replacement of the knee with special reference to prostheses of the condylar variety. In: Galasko CSB, Noble J eds. Current trends in orthopaedic surgery. Manchester University Press.

Freeman MAR, Samuelson KM, Levack B, De Alencar PGC (1986) Knee arthroplasty at the London Hospital 1975-84. Clin Orth 205: 12-20.

Freund DA, Dittus RS, Fitzgerald J, Heck D (1990) Assessing and improving outcomes : total knee replacement. Health Serv Res 25: 723-6.

Fries JF. (1983) The assessment of disability: from first to future principles. Brit J Rheumat 22S: 48-58.

Gartland JJ (1987) Fundamentals of Orthopaedics. 4th edition. Philadelphia: Saunders.

Gill GS, Mills DM (1991) Long term follow up evaluation of 1000 consecutive cemented total knee arthroplasties. Clin Orth 273: 66-76.

Goddard D, Butler R (1984) Rheumatoid arthritis: the treatment controversy. London: Macmillan.

Goddard NJ, Coleman NP (1992) Knee replacement - a new epidemic. In : Joint replacement in the 1990s. The Institute of Mechanical Engineers, Wallace WA, eds. Bury St Edmonds : Mechanical Engineering Publications Ltd.

Goligher JC, Pulvertaft CN, Watkinson G (1964) Controlled trial of vagotomy and gastroenterostomy, vagotomy and antrectomy, and subtotal gastrectomy in elective treatment of duodenal ulcer. Interim report. Br Med J 1: 455-60.

Goodfellow J, O'Connor J (1978) The mechanisms of the knee and prosthesis design. J Bone Joint Surg [Br]; 60-B: 358-69.

Goodfellow JW, O'Connor JJ (1986) Clinical results of the Oxford knee. Surface arthroplasty of the tibiofemoral joint with a meniscal bearing prosthesis. Clin Orthop 205: 21-42.

Goodfellow JW, Kershaw CJ, D'A Benson MK, O'Connor JJ (1988). The Oxford knee for unicompartmental osteoarthritis. J Bone Joint Surg 70-B: 692-701.

Goodfellow J (1993) Science and surgery. J Bone Joint Surg [Br] 75-B: 345-6.

Green SL, Byar DP (1984) The fallacy of omnimetrics. Statistics in Medicine 3: 361-70.

Grewal R, Rimmer RG, Freeman MAR (1992) Early migration of prostheses related to long term survivorship. J Bone Joint Surg [Br] 74-B: 239-42.

Gross M (1993) Innovations in surgery. A proposal for phased clinical trials. J Bone Joint Surg [Br] 75-B: 351-4.

Gunston FH (1971) Polycentric knee arthroplasty: prosthetic simulation of normal knee movement. J Bone Joint Surg [Br] 53-B: 272-7.

Harris WH, Sledge CB (1990) Total hip and total knee replacement. N Engl J Med 323: 801-7.

Hunt SM, McEwen J, McKenna SP (1984) Perceived health: age and sex comparisons in a community. J Epidem Comm Health 38: 156-60.

Insall JN, Ranawat CS, Aglietti P, Shine J (1976) A comparison of four models of total knee replacement prostheses. J Bone Joint Surg [Am] 58-A: 754-61

Insall JN, Binazzi R, Soudry M, Mestriner LA (1985) Total knee arthroplasty. Clin Orthop 192: 13-22.

Insall JN, Dorr LD, Scott RD, Scott WN (1989) Rationale of the knee society clinical rating system. Clin Orthop 248: 13-4.

Jachuck SJ, Brierley H, Jachuck S, Willcox PM (1982) The effect of hypotensive drugs on the quality of life. J Roy Coll Gen Pract 32: 103-5

Jadad AR, MacQuay HJ (1993) The measurement of pain. In: Outcome in orthopaedics (Pynsent PB, Fairbank JCT, Carr A eds) Oxford: Butterworth-Heinemann

Jeffery RS, Morris RW, Denham RA (1991) Coronal alignment after total knee replacement. J Bone Joint Surg [Br] 73-B: 709-14.

Johnson DP, Bannister GC (1986) The outcome of infected arthroplasty of the knee. J Bone Joint Surg 68-B: 289-91.

Kaplan EL, Meier P. (1958) Nonparametric estimation from incomplete observations. J Am Statist Assoc 53: 457-71.

Kellgren JH, Lawrence JS (1963) Atlas of standard radiographs : the epidemiology of chronic rheumatism Vol 2. Oxford : Blackwell

Kirwan JR, Reeback JS (1986). Stanford Health Assessment Questionnaire modified to assess disability in British patients with Rheumatoid Arthritis. Br J Rheumat 25: 266-9.

Kirwan JR, Snow S, Young PJ, Freeman MAR, Currey HLF (1992) Total hip and knee replacement bring long term overall benefit in osteoarthritis and rheumatoid arthritis. Br

J Rheumatol 26S: 78.

Knutson K, Lindstrand A, Lidgren L. (1986) Survival of knee arthroplasties. A nation-wide multicentre investigation of 8000 cases. J Bone Joint Surg [Br] 68-B : 795-803.

Lancet (1986a) The epistemology of surgery. Lancet i : 656-7.

Lancet (1986b) Is there a role for the hinged-knee prosthesis? Lancet ii: 1196-7.

Lancet (1991a) How good are knee replacements? Lancet 338: 477-8.

Lancet (1991b) Informed consent: how informed? Lancet 338: 665-6.

Landis JR, Koch GG. (1977) The measurement of observer agreement for categorical data. Biometrics 33: 159-74.

Laupacis A (1989) The validity of survivorship analysis in total hip arthroplasty (letter). J Bone Joint Surg [Am] 71-A: 1111.

Laupacis A, Rorabeck CH, Bourne RB, Feeny D, Tugwell P, Sim DA (1989) Randomised trials in orthopaedics: why, how and when? J Bone Joint Surg [Am] 71-A: 535-43

Lettin AWF, Neil MJ, Citron ND, August A (1990) Excision arthroplasty for infected constrained total knee replacements. J Bone Joint Surg [Br] 72-B: 220-4.

Lettin AWF, Ware HS, Morris RW (1991) Survivorship analysis and confidence intervals. J Bone Joint Surg [Br] 73-B: 729-31

Liang MH, Cullen KE, Larson MG, Thompson MS, Schwartz JA, Fossel AH, Roberts WN, Sledge CB (1986) Cost-effectiveness of total joint arthroplasty in osteoarthritis.

Arthritis and Rheumatism 29: 937-43.

Love JW (1975) Drugs and operations. Some important differences. JAMA 232: 37-8.

Maquet P (1985) The treatment of choice in osteoarthritis of the knee. Clin Orthop 192 : 108-12.

Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 50: 163-70.

Marmor L (1988) Unicompartmental knee arthroplasty. Ten- to 13- year follow up study. Clin Orthop 228: 14-20.

Meenan RF, Gertman PM, Mason JH (1980) Measuring health status in arthritis: The Arthritis Impact Measurement Scales. Arthritis Rheum 23: 146-52.

Morris RW (1988) A statistical study of papers in the Journal of Bone and Joint Surgery [Br] 1984. J Bone Joint Surg [Br] 70-B: 242-6.

Morris RW (1992a) Survivorship analysis and confidence limits (letter). J Bone Joint Surg [Br] 74-B: 320-1.

Morris RW (1992b) Past and future evaluation of outcome of alternative knee prostheses. J Bone Joint Surg 74-B; Supplement II: 238 (abstract of paper presented at British Association for Surgery of the Knee, May 10 1991)

Morris RW (1993a) Designer hips (letter). Br Med J 306: 1124.

Morris RW (1993b) Bilateral procedures in randomised controlled trials (editorial). J Bone Joint Surg [Br] (in press)

Nellisen RGHH, Brand R, Rozing PM (1992) Survivorship in Total Condylar Knee

Arthroplasty. J Bone Joint Surg [Am] 73-A : 848-57.

Neugebauer E, Troidl H, Spangenberger W, Dietrich A, Lefering R and The Cholycystectomy Study Group (1991) Conventional versus laporoscopic cholycystectomy and the randomized controlled trial. Br J Surg 78: 150-4.

Nilsson KG, Karrholm J, Ekelund L, Magnusson P (1991) Evaluation of micromotion in cemented verus uncemented knee arthroplasty in osteoarthrosis and rheumatoid arthritis. Randomized study using Roentgen sterophotogrammetric Analysis. J Arthroplasty 6: 265-78.

Noble J (1990) Total knee replacement - not a bridge too far. J Bone Joint Surg [Br] 72-B: 173-4 (editorial).

Noble J, Hilton RC (1991) Total knee replacement. Getting better all the time. Br Med J 303: 262.

Oglesby JW, Wilson FC (1984) The evolution of knee arthroplasty. Results with three generations of prostheses. Clin Orth 186 : 96-103

Patrick D (1981) Standardisation of comparative health status measures: using scales developed in America in an English speaking country. Health Service Research Methods: 3rd biennial conference. Hyattsville, Md: US Department of Health and Human Services: 216-29.

Payne CD (ed.) (1986). The GLIM system. Release 3.77. Oxford: Numerical Algorithms Group.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG (1977) Design and analysis of randomised clinical trials requiring prolonged observation of each patient. II Analysis and examples Br J Canc 35 : 1-39.

Pocock SJ (1983) Clinical Trials : A practical approach. Chichester : Wiley.

Pollock AV (1989) The rise and fall of the random controlled trial in surgery. Theor Surg 4: 163-70.

Postel M, Kerboul M, Evrard J, Courpied JP (eds) (1987) Total hip replacement (transl. R. Brueton). Berlin: Springer-Verlag.

Pound P, Gompertz P, Ebrahim S (1992) Can knee replacements be audited by post? (letter) Health Trends 24: 44

Radford PJ (1993) General outcome measures. In: Outcome measures in orthopaedics. (Pynsent PB, Fairbank JCT, Carr A eds.) Oxford: Butterworth-Heinemann.

Ranawat CS, Boachie-Adjei O. (1988) Survivorship analysis and results of total condylar knee arthroplasty. Clin Orthop 226: 6-13.

Rand JA, Ilstrup DM (1991) Survivorship analysis of total knee arthroplasty. J Bone Joint Surg [Am] 73-A: 397-409

Rosner B (1982) Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes. Biometrics 38: 105-14.

Rudicel S, Esdaile J (1985) The randomized clinical trial in orthopaedics : obligation or option? J Bone Joint Surg [Am] 67-A: 1284-93.

Schwartz D, Flamant R, Lellouch J (1980) Clinical Trials (trans. MJR Healy) London : Academic Press

Scuderi GR, Insall JN, Windsor RE, Moran MC (1989) Survivorship of cemented knee replacements. J Bone Joint Surg [Br] 71-B: 798-803.

Sheehan JM (1978) Arthroplasty of the knee. J Bone Joint Surg [Br]; 60-B: 333-8.

Silman AJ (1989) Are there secular trends in the occurrence and severity of rheumatoid arthritis? Scand J Rheumatol Suppl 79: 25-30.

Silman AJ (1991) Is rheumatoid arthritis an infectious disease? An epidemic would answer the question. Br Med J 303: 200.

Snedecor GW, Cochran WG (1980) Statistical methods. Iowa State University Press.

Spector TD (1990) Rheumatoid arthritis. In: Hochberg MC, ed. Epidemiology of rhematic disease. Rheum Dis Clin N Amer 16: 513-37.

Streetly A, Morris R (1992) Can knee replacements be audited by post? (letter) Health Trends 24:74.

Sturrock RD (1991) "Second line" drugs for rheumatoid arthritis. The next generation of drugs needs more rigorous assessment. Br Med J 303: 201.

Sundt T (1987) Was the international trial of extracranial - intercranial bypass representative of the population at risk? N Engl J Med 310: 1363-7.

Tew M, Waugh W (1979) Total replacement of the knee. J Bone Joint Surg 61-B: 225-8.

Tew M, Waugh W (1982) Estimating the survival time of knee replacements. J Bone Joint Surg [Br] 64-B: 579-82.

Tew M, Waugh W (1985) Tibiofemoral alignment and the results of knee replacement. J Bone Joint Surg 67-B: 551-6.

Tew M, Waugh W, Forster IW (1985) Comparing the results of different types of knee

replacement. A method proposed and applied. J Bone Joint Surg 67-B : 775-9.

Thacker C, Fulford P. (1986) Assessment of the Denham knee replacement. Results after eight years. J Bone Joint Surg [Br] 68-B: 600-7.

Thornhill TS, Scott RD (1989) Unicompartmental total knee arthroplasty. Orth Clin North Am 20: 245-56.

Townsend P, Davidson N (1982) Inequalities in health. The Black Report. Penguin

Vandenbroucke JP, Hazevoet HM, Cats A (1984) Survival and cause of death in rheumatoid arthritis: A 25 year prospective follow up. J Rheumatol 11: 158-61.

Walker PS (1989) Requirements for successful total knee replacements. Design considerations. Orth Clin North Amer 20: 15-29.

Wallace WA, Oborne J, Henry DJ (1992) The unmet need for total knee replacement (unpublished)

Waugh W (1978) Knee replacement 1978. J Bone Joint Surg [Br] 60-B : 301-3.

Waugh W (1984) Knee replacement. In : Jackson JP, Waugh W eds. Surgery of the knee joint. London : Chapman and Hall

Weinblatt ME, Maier ML (1991) Disease modifying agents and experimental treatments of rheumatoid arthritis. Clin Orthop 265: 103-15.

Williams A (1991) Setting priorities in health care : an economist's view. J Bone Joint Surg [Br] 73-B : 365-7.

Williams M, Frankel S, Nanchahal K, Coast J, Donovan J (1992) DHA Project: Research Programme. Epidemiologically based needs assessment. Report 2B. Total Knee

Replacement. Bristol : University of Bristol, Health Care Evaluation Unit.

Yates J (1991) Lies, damned lies and waiting list statistics. Br Med J 303: 802.

APPENDIX 1

CHECKLIST FOR ASSESSING STATISTICAL ASPECTS OF SURGICAL PAPERS

Please write comments where necessary

Ring appropriate boxes

Paper number ___

Design of study

1. Is the aim of the study described? Yes / No

2. Is the population properly defined? Yes / No

3. Is the selection of the sample adequately described? Yes / No

4. Is the treatment adequately described? Yes / No

5. Into what scientific category does the study fall?

Case series / Case control study / Series with historical controls / Non-randomised trial with concurrent controls / Randomised controlled trial / Other (please specify)

6. Is there at least one outcome variable which is properly defined?

Yes / No

Results

7. Are the data clearly presented? Yes / No

8. Does the problem require statistical analysis? Yes / No

(If "No", go to question 13)

9. Were any statistical techniques used? Yes / No
(If "No", go to question 13)

10. Are the statistical techniques described? Yes / No

11. Are the statistical techniques appropriate? Yes / No / Unclear

12. Are the calculations correct? Yes / No / Unclear

Please describe below any deficiencies identified in Qs. 9-12

Conclusions

13. Is the title accurate? Yes / No

14. Are the conclusions justified by the results? Yes / No

Extras

15. Is there an explicit mention of a statistician?
Yes, author / Yes, acknowledged / No

16. Number of subjects available for analysis ____

APPENDIX 2

Data from Table I of the publication of Scuderi et al (1989); reproduced with kind permission of the authors and the Journal of Bone & Joint Surgery. "Life table for the original total condylar prosthesis (TCP1)"

	Number				Annual rates		
Follow up time (mths)	No. at start	No. who fail	No. who with-drew	No. at risk	Fail rate	Success rate	Overall success
0 - 12	224	1	28	210	0.48	99.52	99.52
13 - 24	195	2	16	187	1.07	98.93	98.46
25 - 36	177	1	19	167.5	0.60	99.40	97.86
37 - 48	157	0	7	153.5	0	100	97.86
49 - 60	150	0	18	141	0	100	97.86
61 - 72	132	2	23	120.5	1.66	98.34	96.24
73 - 84	107	4	3	105.5	3.79	96.21	92.59
85 - 96	100	0	6	97	0	100	92.59
97- 108	94	2	5	91.5	2.18	97.81	90.56
109-120	87	0	16	79	0	100	90.56
121-132	71	0	20	61	0	100	90.56
133-144	51	0	13	44.5	0	100	90.56
144-156	38	0	28	24	0	100	90.56
157-168	10	0	9	5.5	0	100	90.56
169-180	1	0	1	0.5	0	100	90.56

APPENDIX 3. Relationship between pain measures at given time points and various numbers of time periods hence.

In all tables, missed follow ups are included with the missing category

Table A3(i). Relationship between pain measures at given time point and pain measure two periods hence.

Pain during activity

(n+2) th follow up

nth follow up

	None	Mild	Mod/Sev	Missing	Total
None	287	68	17	345	717
	(40%)	(9%)	(2%)	(48%)	(100%)
Mild	70	56	16	144	286
	(24%)	(20%)	(6%)	(50%)	(100%)
Mod/Sev	19	17	22	49	107
	(18%)	(16%)	(21%)	(46%)	(100%)
Missing	143	55	21	1219	1438
	(10%)	(4%)	(1%)	(85%)	(100%)
Total	519	196	76	1757	2548

Table A3(ii). Relationship between pain measures at given time point and pain measure three periods hence.

Pain during activity

(n+3) th follow up

nth follow up

	None	Mild	Mod/Sev	Missing	Total
None	203	57	12	397	669
	(30%)	(9%)	(2%)	(59%)	(100%)
Mild	55	41	8	158	262
	(21%)	(16%)	(3%)	(60%)	(100%)
Mod/Sev	15	12	19	57	103
	(15%)	(12%)	(18%)	(55%)	(100%)
Missing	111	37	15	987	1150
	(10%)	(3%)	(1%)	(86%)	(100%)
Total	384	147	54	1599	2184

Table A3 (iii). Relationship between pain measures at given time point and pain measure four periods hence.

Pain during activity

(n+4) th follow up

nth follow up

	None	Mild	Mod/Sev	Missing	Total
None	143	43	11	402	599
	(24%)	(7%)	(2%)	(67%)	(100%)
Mild	35	24	7	167	233
	(15%)	(10%)	(3%)	(72%)	(100%)
Mod/Sev	11	6	9	65	91
	(12%)	(7%)	(10%)	(71%)	(100%)
Missing	80	28	15	774	897
	(9%)	(3%)	(2%)	(86%)	(100%)
Total	269	101	42	1408	1820

Table A3 (iv). Relationship between pain measures at given time point and pain measure five periods hence.

Pain during activity

(n+5) th follow up

nth follow up

	None	Mild	Mod/Sev	Missing	Total
None	83	22	9	397	511
	(16%)	(4%)	(2%)	(78%)	(100%)
Mild	28	18	5	149	200
	(14%)	(9%)	(3%)	(75%)	(100%)
Mod/Sev	8	6	6	58	78
	(10%)	(8%)	(8%)	(74%)	(100%)
Missing	62	22	9	574	667
	(9%)	(3%)	(1%)	(86%)	(100%)
Total	181	68	29	1178	1456

APPENDIX 4. Bias in estimation of survival probabilities caused by assuming two knee replacements carried out on a single patient do not differ in their survival from single knee replacements.

Two probabilities are calculated, namely that of 5 year survival for patients' first knee operation (p1) and that of all knees (p2).

S will denote survival at 5 years
U will denote a patient with one knee replaced
B will denote a patient with both knees replaced
B1 will denote the first of two knees replaced
B2 will denote the second of two knees replaced.

pr(A) denotes the probability of an event A occurring

$$p1 = pr(S \mid U) \times pr(U) + pr(S \mid B1) \times pr(B)$$

$$p2 = \frac{[\, pr(S \mid U) \times pr(U) + pr(S \mid B1) \times pr(B) + pr(S \mid B2) \times pr(B)]}{pr(U) + 2 \times pr(B)}$$

Observed data in Chapter 5 suggest:

$pr(U) = 0.726$, so $pr(B) = 0.274$

$pr(S \mid U) = 0.8163$, $pr(S \mid B1) = 0.8874$, $pr(S \mid B2) = 0.9095$

Therefore $p1 = 0.836$ and $p2 = 0.852$

The table below shows differences between p1 and p2 under different scenarios

pr(U)	pr(S|U)	pr(S|B1)	pr(S|B2)	p1	p2
0.1	0.8	0.9	0.9	0.81	0.818
0.25	0.8	0.9	0.9	0.825	0.84
0.5	0.8	0.9	0.9	0.85	0.867
0.5	0.7	0.9	0.9	0.8	0.833

The extent of overestimation is seen to depend not so much on pr(U), the proportion of unilaterally replaced patients in the series, but on the size of the difference between pr(S|U) on the one hand (survival amongst "unilaterals") and pr(S|B1) and pr(S|B2) (survival amongst "bilaterals") on the other.

Appendix 5

QUESTIONNAIRE TO PATIENTS WHO HAVE UNDERGONE TOTAL KNEE REPLACEMENT ☐☐☐☐☐

We are interested to find out how your knee replacement affects everyday activities. For each question, please tick the box which best describes your opinion.

1. Can you dress yourself?
   - Not without someone's help ☐ 1
   - By myself with difficulty ☐ 2
   - By myself without difficulty ☐ 3

2. Can you stand up from a straight chair?
   - Not without someone's help ☐ 1
   - By myself if I hold on ☐ 2
   - By myself without holding on ☐ 3

3. Can you go upstairs?
   - I never use stairs ☐ 1
   - Only with someone's help ☐ 2
   - I usually hold the bannister and lead with the same leg ☐ 3
   - I usually hold the bannister and lead with alternate legs ☐ 4
   - I can go upstairs without holding at all ☐ 5

4. Can you go downstairs?
   - I never use stairs ☐ 1
   - Only with someone's help ☐ 2
   - Only if I face backwards ☐ 3
   - I usually hold the bannister and lead with the same leg ☐ 4
   - I usually hold the bannister and lead with alternate legs ☐ 5
   - I can go downstairs without holding at all ☐ 6

5. For how long can you walk at one time?
   - Not at all ☐ 1
   - Only indoors ☐ 2
   - For up to 10 minutes ☐ 3
   - For up to 30 minutes ☐ 4
   - For up to 60 minutes ☐ 5
   - For more than 60 minutes ☐ 6

6. How do you get out of bed?
   - Not without someone's help ☐ 1
   - Without help if I hold on ☐ 2
   - Without help and without holding on ☐ 3

7. Can you pick clothing off the floor?
   - No problem ☐ 1
   - With a bit of difficulty ☐ 2
   - With a lot of difficulty ☐ 3
   - Impossible ☐ 4

8. Can you do household chores like vaccuum cleaning?
   - No problem ☐ 1
   - With a bit of difficulty ☐ 2
   - With a lot of difficulty ☐ 3
   - Impossible ☐ 4

9. Can you cut your own toe-nails?

No problem	☐ 1
With a bit of difficulty	☐ 2
With a lot of difficulty	☐ 3
Impossible	☐ 4

10.Do you use any special devices? (tick one box for each device)

	Never 1	Sometimes 2	Often 3
(a) Walking cane	☐	☐	☐
(b) Crutches	☐	☐	☐
(c) Walking frame	☐	☐	☐

11. Which of the following joints have you had replaced?(tick one box for each joint)

	No 1	Yes 2
(a) Right knee	☐	☐
(b) Left knee	☐	☐
(c) Right hip	☐	☐
(d) Left hip	☐	☐
(e) Other joint	☐	☐

12. Have you experienced pain in the past week in any of the following joints? (tick one box for each joint)

	No 1	Only whilst moving 2	Whilst at rest 3
(a) Right knee	☐	☐	☐
(b) Left knee	☐	☐	☐
(c) Right hip	☐	☐	☐
(d) Left hip	☐	☐	☐
(e) Right ankle or foot	☐	☐	☐
(f) Left ankle or foot	☐	☐	☐

Please fill in the following questions only if you have had your **RIGHT** knee replaced. (Otherwise skip this page and the next one)

Please answer each question by ringing one of the seven marks on the line below the question. For instance in the first question, you would ring "0" if you felt no better compared with before the operation, "1" if you felt just slightly better, "2" if you felt a bit better, and so on.

1.How do you feel about this knee now compared with before the operation?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
No better — Very much better

2. Has this knee improved as much as you expected before the operation?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
Much less than expected — Much more than expected

3. How do you feel about this knee compared with one year ago?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
Much worse — Much better

4. How much difference has this knee replacement made to your overall health?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
No difference — A huge difference

5. How worthwhile do you feel that this knee replacement has been?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
Not worth the trouble — Very worthwhile

6a. Whilst resting, how much pain have you had from this knee in the past week?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
No pain — Very severe pain

6b. Whilst moving around, how much pain have you had from this knee in the past week?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
No pain — Very severe pain

7. Are you able to bend this knee as much as you would wish?

No ☐ 1
Yes ☐ 2

8. Can you sit with this knee bent in a confined space (e.g in the cinema or theatre)

Not at all ☐ 1
For up to 10 minutes ☐ 2
For up to 30 minutes ☐ 3
For more than 30 minutes ☐ 4

Please fill in the following questions only if you have had your **LEFT** knee replaced. (Otherwise skip this page and the next one)

Please answer each question by ringing one of the seven marks on the line below the question. For instance in the first question, you would ring "0" if you felt no better compared with before the operation, "1" if you felt just slightly better, "2" if you felt a bit better, and so on.

1.How do you feel about this knee now compared with before the operation?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
No better — Very much better

2. Has this knee improved as much as you expected before the operation?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
Much less than expected — Much more than expected

3. How do you feel about this knee compared with one year ago?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
Much worse — Much better

4. How much difference has this knee replacement made to your overall health?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
No difference — A huge difference

5. How worthwhile do you feel that this knee replacement has been?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
Not worth the trouble — Very worthwhile

6a. Whilst resting, how much pain have you had from this knee in the past week?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
No pain — Very severe pain

6b. Whilst moving around, how much pain have you had from this knee in the past week?

0 1 2 3 4 5 6
|_____|_____|_____|_____|_____|_____|
No pain — Very severe pain

7. Are you able to bend this knee as much as you would wish?

No ☐ 1
Yes ☐ 2

8. Can you sit with this knee bent in a confined space (e.g in the cinema or theatre)

Not at all ☐ 1
For up to 10 minutes ☐ 2
For up to 30 minutes ☐ 3
For more than 30 minutes ☐ 4

Now I would like to ask a few questions about your home and family.

1. What is your marital status? (tick one answer)

Married	☐ 1
Single	☐ 2
Widowed	☐ 3
Divorced/Separated	☐ 4
Other	☐ 5

2. What sort of home do you have? (tick one answer)

Rented from council	☐ 1
Rented privately	☐ 2
Part of housing association	☐ 3
Owned by you or your family, with or without mortgage	☐ 4
Old people's home	☐ 5

3. Which people live with you? (tick one box for each person listed)

	No 1	Yes 2
(a) Husband/wife	☐	☐
(b) Brother/sister	☐	☐
(c) Son/daughter	☐	☐
(d) Other relation	☐	☐
(e) Friend	☐	☐

Thank you very much for your time and effort in answering these questions. We are interested in any other comments you may have about your knee replacement and about this questionnaire:-

Appendix 6

UNITED MEDICAL AND DENTAL SCHOOLS
OF
GUY'S AND ST. THOMAS'S HOSPITALS
(UNIVERSITY OF LONDON)

DIVISION OF COMMUNITY HEALTH

Professor W W Holland
Fax No. 01-928 1468
Tel. 01-928 9292 ext.

Correspondence to:
DEPARTMENT OF COMMUNITY MEDICINE
ST. THOMAS'S CAMPUS, LONDON SE1 7EH

Dear Sir/Madam

STUDY OF OUTCOME OF TOTAL KNEE REPLACEMENT

We would very much appreciate your help with a study of people who have undergone a total knee replacement operation. We are comparing the feelings of people who have received different types of knee replacement. This survey is being run by the Orthopaedic Academic Unit and the Department of Community Medicine of St Thomas's Hospital, London. Your reply will be completely confidential and will only be used for research purposes. It will not be seen by Mr Heatley or any other surgeons involved in caring for you.

Could you therefore fill in the enclosed questionnaire as soon as possible and return it to us in the pre-paid addressed envelope? The information you give us will be helpful in answering some important medical questions.

Yours sincerely

Mr F W Heatley
(Consultant Orthopaedic Surgeon)

Mr Richard Morris
(Research Co-ordinator)

UNITED MEDICAL AND DENTAL SCHOOLS
OF
GUY'S AND ST. THOMAS'S HOSPITALS
(UNIVERSITY OF LONDON)

DIVISION OF COMMUNITY HEALTH

Professor W W Holland
Fax No. 01-928 1468
Tel. 01-928 9292 ext.

Correspondence to:
DEPARTMENT OF COMMUNITY MEDICINE
ST. THOMAS'S CAMPUS, LONDON SE1 7EH

Dear Sir/Madam,

Study of outcome of Total Knee Replacement

You should have already received a letter from Mr. Heatley and myself asking you to fill in a questionnaire about your knee replacement. Unfortunately we have not received some questionnaires back. If you have already sent back your questionnaire, please ignore this reminder.

If you have not returned your questionnaire, we would be very grateful if you could now fill it in and return it to us WITHIN TWO DAYS in the prepaid envelope which we sent you with the questionnaire. It is very important that we have the views of all patients who have undergone a knee replacement operation. It should only take a few minutes of your time, as most questions only need a tick or a ring against the answer which applies to you. Your answers will be kept confidential, and will not be seen by Mr. Heatley or other surgeons involved in caring for you.

Thank you for your co-operation in this survey. The information will be very helpful in understanding how successful knee replacement operations are.

Yours sincerely,

Richard Morris

Richard Morris
(Research co-ordinator)

UNITED MEDICAL AND DENTAL SCHOOLS
OF
GUY'S AND ST.THOMAS'S HOSPITALS
(UNIVERSITY OF LONDON)

DEPARTMENT OF COMMUNITY MEDICINE

PROFESSOR J.A D. ANDERSON

GUY'S HOSPITAL
MEDICAL SCHOOL BUILDING
LONDON BRIDGE, SE1 9RT
TELEPHONE 01-407 7600 Ext. 2157

22nd September 1989

Dear

About six weeks ago, a questionnaire was sent to you by myself and Mr. F. Heatley of St. Thomas' Hospital concerning your knee replacement. According to our records, we have not yet received a reply from you. We do wish to hear from you because it is important to us that we have the views of all people who have undergone a knee replacement operation.

If you have not yet returned a questionnaire to us, we would be very grateful if you could fill in the enclosed questionnaire. It should be completed WITHIN 2 DAYS and returned to us in the enclosed pre-paid envelope. It should only take a few minutes of your time, as most questions only need a tick or ring against the answer which applies to you. Your answers will be kept confidential and will not be seen by Mr. Heatley or other surgeons involved in caring for you.

Thank you for your co-operation. The information you give will be very helpful in understanding how successful knee replacement operations are.

Yours sincerely,

Richard Morris
(Research Co-ordinator, based at Guy's Hospital).

APPENDIX 7. Weighting of knee scores from patients with both knees replaced

To assess the relative weights to be given in analysis of knee scores to patients with one knee replaced only (unilateral) and those with both knees replaced (bilateral), a special preliminary analysis was carried out. It was assumed in the main analysis that bilateral patients would have the average of their two knee scores used.

Material

Three hundred and fifty nine patients with both knees replaced provided two knee scores. Components of variance (between subject and within subject) were calculated by the standard method described by Armitage and Berry (1987). This gave estimates as follows:

Between subject variance, $\sigma_B^2 = 73.3$

Within subject variance, $\sigma_W^2 = 28.4$

This gives an intraclass correlation of 0.72. It was recommended by Rosner (1982) that regarding multiple observations on the same subject as statistically independent was to be avoided if the intraclass correlation exceeded 0.4.

Method

Let x = Knee score for a unilateral patient

Let x_1, x_2 = Knee scores for a bilateral patient

Thus $(x_1 + x_2)/2$ = Average knee score for a bilateral patient

Now Variance(x) = $\sigma_B^2 + \sigma_W^2$,

The usual convention is that weights are inversely proportional to the variance (Armitage and Berry 1987, pages 194-6).

Thus $w_{uni} = k\,[1/(\sigma_B^2 + \sigma_W^2)]$ - (1)

and $w_{bil} = k\,[1/(\sigma_B^2 + \sigma_W^2/2]$ - (2)

In all there are 819 observations available for analysis of the knee score: 359 bilaterals and 460 unilaterals.

We require that $460 \times w_{uni} + 359 \times w_{bil} = 819$ - (3)

Rearranging (1), (2), (3) gives:

$w_{uni} = 0.933791$ and $w_{bil} = 1.08477$

Remarks

Although the knee score was dichotomised for formal analysis and the current weights are designed for analysis of knee score as a continuous variable, the above strategy was thought a reasonable approximation.

The analysis was carried out both with the above weights and with equal weights for every patient, whether unilateral or bilateral. The difference in estimated odds ratios according to the two analyses was an order of magnitude less than the estimates themselves, and the difference in standard errors was two orders of magnitude less than the standard errors themselves. Examples are shown below:

Factor		Log Odds ratio (standard error)	
Subgroup	Compared with	Equally weighted	Weighted for bilaterals
Surgeon series			
Series 09	Series 01	0.84 (0.47)	0.89 (0.47)
Time since operation			
Under 3 yrs	Over 10 yrs	-0.95 (0.35)	-0.96 (0.35)
Gender			
Females	Males	-0.21 (0.19)	-0.22 (0.19)

Appendix 8

QUESTIONNAIRE TO PATIENTS WHO HAVE UNDERGONE TOTAL KNEE REPLACEMENT

We are interested to find out how your knee replacement affects everyday activities. For each question, please tick the box which best describes your opinion.

N			%
1. Can you dress yourself?			
149	Not without someone's help	1	16
279	By myself with difficulty	2	31
471	By myself without difficulty	3	52
8	NK		1
2. Can you stand up from a straight chair?			
97	Not without someone's help	1	11
518	By myself if I hold on	2	57
283	By myself without holding on	3	31
9	NK		1
3. Can you go upstairs?			
198	I never use stairs	1	22
36	Only with someone's help	2	4
392	I usually hold the bannister and lead with the same leg	3	43
209	I usually hold the bannister and lead with alternate legs	4	23
52	I can go upstairs without holding at all	5	6
20	NK		2
4. Can you go downstairs?			
190	I never use stairs	1	21
31	Only with someone's help	2	3
83	Only if I face backwards	3	9
394	I usually hold the bannister and lead with the same leg	4	43
153	I usually hold the bannister and lead with alternate legs	5	17
31	I can go downstairs without holding at all	6	3
25	NK		3
5. For how long can you walk at one time?			
28	Not at all	1	3
137	Only indoors	2	15
218	For up to 10 minutes	3	24
235	For up to 30 minutes	4	26
130	For up to 60 minutes	5	14
136	For more than 60 minutes	6	15
23	NK		3
6. How do you get out of bed?			
55	Not without someone's help	1	6
435	Without help if I hold on	2	48
409	Without help and without holding on	3	45
8	NK		1
7. Can you pick clothing off the floor?			
392	No problem	1	43
323	With a bit of difficulty	2	36
115	With a lot of difficulty	3	13
67	Impossible	4	7
10	NK		1
8. Can you do household chores like vaccuum cleaning?			
346	No problem	1	38
234	With a bit of difficulty	2	26
124	With a lot of difficulty	3	14
189	Impossible	4	21
14	NK		1

9. Can you cut your own toe nails?

117	No problem	☐ 1	13
192	With a bit of difficulty	☐ 2	21
130	With a lot of difficulty	☐ 3	14
415	Impossible	☐ 4	46
53	NK		6

10.Do you use any special devices? (tick one box for each device)

	Never 1	Sometimes 2	Often 3	NK
(a) Walking cane	308 (34%)	172 (19%)	357 (39%)	70 (8%)
(b) Crutches	758 (84%)	26 (3%)	46 (5%)	77 (8%)
(c) Walking frame	753 (83%)	38 (4%)	46 (5%)	70 (8%)

11. Which of the following joints have you had replaced?(tick one box for each joint)

	No 1	Yes 2	NK
(a) Right knee	223 (25%)	678 (75%)	6 (1%)
(b) Left knee	272 (30%)	625 (69%)	10 (1%)
(c) Right hip	673 (74%)	175 (19%)	59 (7%)
(d) Left hip	746 (82%)	133 (15%)	28 (3%)
(e) Other joint	813 (90%)	76 (8%)	18 (2%)

12. Have you experienced pain in the past week in any of the following joints? (tick one box for each joint)

	No 1	Only whilst moving 2	Whilst at rest 3	NK
(a) Right knee	332 (37%)	253 (28%)	144 (16%)	178 (20%)
(b) Left knee	337 (37%)	234 (26%)	121 (13%)	215 (24%)
(c) Right hip	367 (40%)	95 (10%)	52 (6%)	393 (43%)
(d) Left hip	383 (42%)	63 (7%)	41 (5%)	420 (46%)
(e) Right ankle or foot	306 (34%)	141 (16%)	75 (8%)	385 (42%)
(f) Left ankle or foot	328 (36%)	131 (14%)	70 (8%)	378 (42%)

Please fill in the following questions only if you have had your **RIGHT** knee replaced. (Otherwise skip this page and the next one)

Please answer each question by ringing one of the seven marks on the line below the question. For instance in the first question, you would ring "0" if you felt no better compared with before the operation, "1" if you felt just slightly better, "2" if you felt a bit better, and so on.

1.How do you feel about this knee now compared with before the operation?

%	5	3	4	9	7	16	58
	0	1	2	3	4	5	6
	No better						Very much better

N = 673

2. Has this knee improved as much as you expected before the operation?

%	7	4	5	7	9	21	47
	0	1	2	3	4	5	6
	Much less than expected						Much more than expected

N = 669

3. How do you feel about this knee compared with one year ago?

%	6	5	4	18	10	16	40
	0	1	2	3	4	5	6
	Much worse						Much better

N = 657

4. How much difference has this knee replacement made to your overall health?

%	12	5	3	7	9	16	48
	0	1	2	3	4	5	6
	No difference						A huge difference

N = 660

5. How worthwhile do you feel that this knee replacement has been?

%	4	3	2	5	4	14	69
	0	1	2	3	4	5	6
	Not worth the trouble						Very worthwhile

N = 671

6a. Whilst resting, how much pain have you had from this knee in the past week?

%	49	18	11	10	5	3	5
	0	1	2	3	4	5	6
	No pain						Very severe pain

N = 670

6b. Whilst moving around, how much pain have you had from this knee in the past week?

%	38	17	13	12	10	4	6
	0	1	2	3	4	5	6
	No pain						Very severe pain

N = 665

N		%	
7. Are you able to bend this knee as much as you would wish?			
314	No	49	☐ 1
326	Yes	51	☐ 2

8. Can you sit with this knee bent in a confined space (e.g in the cinema or theatre)

124	Not at all	19	☐ 1
89	For up to 10 minutes	14	☐ 2
154	For up to 30 minutes	24	☐ 3
271	For more than 30 minutes	43	☐ 4

Please fill in the following questions only if you have had your **LEFT** knee replaced. (Otherwise skip this page and the next one)

Please answer each question by ringing one of the seven marks on the line below the question. For instance in the first question, you would ring "0" if you felt no better compared with before the operation, "1" if you felt just slightly better, "2" if you felt a bit better, and so on.

1.How do you feel about this knee now compared with before the operation?

% 7 2 5 8 8 13 57

0 1 2 3 4 5 6

No better — Very much better

N = 616

2. Has this knee improved as much as you expected before the operation?

% 9 4 5 7 11 18 46

0 1 2 3 4 5 6

Much less than expected — Much more than expected

N = 613

3. How do you feel about this knee compared with one year ago?

% 6 5 7 19 10 13 41

0 1 2 3 4 5 6

Much worse — Much better

N = 595

4. How much difference has this knee replacement made to your overall health?

% 13 4 2 7 10 16 48

0 1 2 3 4 5 6

No difference — A huge difference

N = 598

5. How worthwhile do you feel that this knee replacement has been?

% 5 3 3 5 4 13 67

0 1 2 3 4 5 6

Not worth the trouble — Very worthwhile

N = 611

6a. Whilst resting, how much pain have you had from this knee in the past week?

% 50 17 9 9 7 3 5

0 1 2 3 4 5 6

No pain — Very severe pain

N = 610

6b. Whilst moving around, how much pain have you had from this knee in the past week?

% 40 17 11 11 10 4 7

0 1 2 3 4 5 6

No pain — Very severe pain

N = 610

N		%	
7. Are you able to bend this knee as much as you would wish?			
261	No	45	☐ 1
318	Yes	55	☐ 2

8. Can you sit with this knee bent in a confined space (e.g in the cinema or theatre)

112	Not at all	19	☐ 1
82	For up to 10 minutes	14	☐ 2
125	For up to 30 minutes	22	☐ 3
259	For more than 30 minutes	45	☐ 4

Now I would like to ask a few questions about your home and family.

N		%	
	1. What is your marital status? (tick one answer)		
466	Married	51	☐ 1
72	Single	8	☐ 2
308	Widowed	34	☐ 3
37	Divorced/Separated	4	☐ 4
0	Other	0	☐ 5
24	NK	3	
	2. What sort of home do you have? (tick one answer)		
213	Rented from council	24	☐ 1
49	Rented privately	5	☐ 2
34	Part of housing association	4	☐ 3
549	Owned by you or your family, with or without mortgage	61	☐ 4
23	Old people's home	3	☐ 5
39	NK	4	

3. Which people live with you? (tick one box for each person listed)

	No 1	Yes 2	NK
(a) Husband/wife	420 (47%)	461 (51%)	26 (3%)
(b) Brother/sister	578 (64%)	21 (2%)	308 (34%)
(c) Son/daughter	516 (57%)	143 (16%)	248 (27%)
(d) Other relation	576 (64%)	22 (2%)	309 (34%)
(e) Friend	573 (63%)	22 (2%)	312 (34%)

Thank you very much for your time and effort in answering these questions. We are interested in any other comments you may have about your knee replacement and about this questionnaire:-

APPENDIX 9. Comparison of various items of questionnaire with surgeons' assessments at recent follow up clinics

Throughout this appendix, * shows cells of the table representing agreement between surgeon and patient.

Table A9(i). Comparison of surgeons' and patients' assessments of rising from a straight chair (Series 04/06)

	Patient		
Surgeon	Need help	If hold on	No holding
Unable	2*	0	0
If hold on	2	9*	4
Difficult	0	6	2*
Easy	0	2	5*

Table A9(ii). Comparison of surgeons' and patients' assessments of climbing stairs (Series 04/06)

	Patient				
Surgeon	Never use	Need help	Bann.- same leg	Bann.- alternate legs	No hold
Bizarre/ unable	4*	1	0	1	0
Bannister	1	0*	5	1	0
One step	2	0	7*	3	1
Normal	0	0	1	4*	1*

Table A9(iii). Comparison of surgeons' and patients' assessments of descending stairs (Series 04/06)

	Patient					
Surgeon	Never use	Need help	Face back	Bann.- lead same leg	Bann.- lead altern. leg	No hold
Bizarre	4*	1*	0*	0	2	0
Bann.	1	0	1	6*	0	0
One step	2	0	3	6*	4	0
Normal	0	0	0	1	1*	0*

Table A9(iv). Comparison of surgeons' and patients' assessments of walking distance (Series 04/06)

	Patient					
Surgeon	Not at all	Indoors only	Up to 10 mins	10-30 mins	30-60 mins	Over 60 mins
Unable	1*	0	0	0	0	0
Indoors only	0	3*	1	0	0	0
Up to 10 mins	0	1	3*	2	2	0
10-30 mins	0	1	2	2*	0	0
30-60 mins	0	0	0	2	1*	0
Over 60 mins	0	0	0	0	3	2*
Unlimited	0	0	0	2	0	4*

Table A9(v). Comparison of surgeons' and patients' assessments of walking stick usage (Series 04/06)

	Patient		
Surgeon	Never	Sometimes	Often
None	7*	2	4
Stick outside	0	2*	4
2 sticks outside	0	0*	1
Stick always	1	2	1*
2 sticks always	0	0	1*
Crutches	0	0	0
Unable	0	0	0

Table A9(vi). Comparison of surgeons' and patients' assessments of pain in activity (knees) (Series 04/06)

Patient

Whilst moving around, how much pain have you had from this knee in the past week?

Surgeon

Severity of pain	0	1	2	3	4	5	6
None	21*	8	3	1	4	0	0
Mild	0	0*	1*	2	1	0	0
Significant	3	0	0	0*	3*	0	0
Severe	0	0	0	0	0	0*	0*

Table A9(vii). Comparison of surgeons' and patients' assessments of worthwhileness of operation (no. knees) (Series 04/06)

Patient

How worthwhile do you feel that this knee replacement has been?

Surgeon

With hindsight, would patient have operation again?	0	1	2	3	4	5	6
No	2*	0*	1	0	0	0	6
Unsure	0	0	0*	0*	0*	0	0
Yes	0	0	0	2	0	8*	26*

Table A9(viii). Comparison of surgeons' and patients' assessments of rising from a straight chair (Series 05)

		Patient	
Surgeon	Need help	If hold on	No holding
Unable	0*	2	0
If hold on	1	26*	2
Difficult	2	13	6*
Easy	0	5	3*

Appendix A9(ix). Comparison of surgeon's and patients' assessments of climbing stairs (Series 05)

			Patient		
Surgeon	Never use	Need help	Bann.- same leg	Bann.- alternate legs	No hold
Bizarre/ unable	7*	0*	2	0	0
Bannister	7	0	23*	7*	2
No rail	3	0	4	3	0*

Table A9(x). Comparison of surgeon's and patients' assessments of descending stairs (Series 05)

	Patient					
Surgeon	Never use	Need help	Face back	Bann.-same leg	Bann.-altern. leg	No hold
Unable	7*	1*	0*	1	0	0
Bannister	7	0	9	21*	3*	0
Normal	3	0	1	6	0	0*

Table A9(xi). Comparison of surgeon's and patients' assessments of walking ability (Series 05)

	Patient					
Surgeon	Not at all	Indoors only	Up to 10 mins	10-30 mins	30-60 mins	Over 60 mins
Unable	0*	0	1	0	0	0
Housebound	0	4*	3	0	0	0
<400 m	0	2	7*	3*	0	0
400-800 m	0	2	6	2	0*	1
>800 m	0	1	0	5	3	2*
Unlimited	0	1	1	4	5	6*

Table A9(xii). Comparison of surgeon's and patients' assessments of use of walking sticks (Series 05)

		Patient	
Surgeon	Never	Sometimes	Often
No stick	12*	6	12
1 stick	0	4*	16*
2 sticks	0	1*	0*
Crutches/ frame	0	0*	2*

Table A9(xiii). Comparison of surgeon's and patients' assessments of resting pain (Series 05)

	Patient Whilst resting, how much pain have you had from this knee in the past week?						
Surgeon	0	1	2	3	4	5	6
None	45*	10	5	7	2	0	0
Mild	0	0*	2*	1	3	0	1
Moderate	0	0	4	2*	1*	0	0
Severe	0	0	0	0	0	1*	0*

Table A9(xiv). Comparison of surgeon's and patients' assessments of moving pain (Series 05)

	Patient						
	Whilst moving around, how much pain have you had from this knee in the past week?						
Surgeon	0	1	2	3	4	5	6
None	34*	3	4	4	5	0	0
Mild	5	3*	3*	2	5	0	1
Moderate	1	0	3	2*	3*	1	1
Severe	0	0	0	2	0	1*	1*

Table A9(xv). Comparison of surgeons' and patients' assessments of rising from a straight chair (Series 07)

		Patient	
Surgeon	Need help	If hold on	No holding
Unable	1*	0	0
If hold on	2	3*	12
Difficult	2	2	7
Easy	3	2	42*

Appendix A9(xvi). Comparison of surgeon's and patients' assessments of climbing stairs (Series 07)

	Patient				
Surgeon	Never use	Need help	Bann.- same leg	Bann.- alternate legs	No hold
Otherwise	4*	1*	0	0	0
Bannister only	2	0	6*	5	0
One at a time	2	1	12*	10*	0
No rail	2	0	7	13	12*

Table A9(xvii). Comparison of surgeon's and patients' assessments of walking ability (Series 07)

	Patient					
Surgeon	Not at all	Indoors only	Up to 10 mins	10-30 mins	30-60 mins	Over 60 mins
Indoors only	0	0*	0	0	0	0
5-10 mins	0	3	4*	2	1	1
10-30 mins	0	1	4	4*	4	0
30-60 mins	0	0	0	11	10*	8
>60 mins	0	0	3	12	3	12*
Can run	0	0	1	0	1	1*

Table A9(xviii). Comparison of surgeons' and patients' assessments of worthwhileness of operation (no. knees) (Series 07)

Patient

How worthwhile do you feel that this knee replacement has been?

Surgeon

	0	1	2	3	4	5	6
Disappointed	0*	0*	0	0	0	0	3
Non-committal	0	0	0*	0*	1	0	1
Satisfied	0	0	0	2	0*	4*	18
Enthusiastic	0	0	0	1	2	6	24*

Table A9(xix). Comparison of surgeon's and patients' assessments of resting pain (Series 07)

Patient

Whilst resting, how much pain have you had from this knee in the past week?

Surgeon	0	1	2	3	4	5	6
None	34*	17	4	5	1	3	0
Mild	5	1*	1*	2	1	1	1
Moderate	1	0	0	0*	1*	1	0
Severe	0	0	0	0	0	0*	0*

Table A9(xx). Comparison of surgeon's and patients' assessments of moving pain (Series 07)

	Patient Whilst moving around, how much pain have you had from this knee in the past week?						
Surgeon	0	1	2	3	4	5	6
None	23*	13	4	4	1	2	0
Mild	5	5*	3*	3	3	3	2
Moderate	0	1	0	1*	0*	2	0
Severe	0	0	0	0	0	0*	0*

Table A9(xxi). Comparison of surgeon's and patients' assessments of resting pain (Series 08)

	Patient Whilst resting, how much pain have you had from this knee in the past week?						
Surgeon	0	1	2	3	4	5	6
None	11*	11	3	3	0	0	0
Mild	2	3*	1*	2	0	0	1
Moderate remittant	0	0	0	0*	0	0	0
Moderate constant	0	0	0	0	0*	0	0
Severe remittant	0	0	0	0	0	0*	0
Severe constant	0	0	0	0	0	0	0*

Table A9(xxii). Comparison of surgeon's and patients' assessments of pain in activity (Series 08)

	Patient Whilst moving around, how much pain have you had from this knee in the past week?						
Surgeon	0	1	2	3	4	5	6
None	15*	7	3	3	0	0	0
Mild	4	1*	2*	1	0	1	0
Moderate remittant	0	0	0	0*	0	0	0
Moderate constant	0	0	0	0	0*	0	0
Severe remittant	0	0	0	0	0	0*	0
Severe constant	0	0	0	0	0	0	0*

Table A9(xxiii). Comparison of surgeons' and patients' assessments of worthwhileness of operation (no. knees) (Series 08)

Patient

How worthwhile do you feel that this knee replacement has been?

Surgeon

	0	1	2	3	4	5	6
Poor	0*	0*	0	0	0	0	0
Fair	0	0	0*	0*	0	0	1
Good	0	0	0	1	0*	1*	2
Excellent	1	0	0	1	1	7	23*

Table A9(xxiv). Comparison of "Doctor's assessment" with patients' assessments of the knee compared with before the operation (Series 08)

Patient

How do you feel about this knee now compared with before the operation?

Surgeon

	0	1	2	3	4	5	6
Poor	0*	0*	0	0	0	0	0
Fair	0	0	0*	0*	0	0	1
Good	0	0	0	1	0*	2*	2
Excellent	0	1	1	2	1	5	22*

APPENDIX 10

According to the analysis of Chapter 10, the knee score was dichotomised at its overall median so that half the patients would exceed it. Thus the variance of the difference in proportion (p) of good outcomes between two groups of 525 patients, using the formula for a Binomial variance, would be

$$p\,(1 - p)\left(\frac{1}{525} + \frac{1}{525}\right) = 9.5238 \times 10^{-4} \text{ assuming } p=0.5 \quad (1)$$

We may estimate the surgeon effect for prostheses A, D and E separately since two surgeons used each of these prostheses.

Prosthesis A; proportions of good outcomes were 0.5 and 0.464 for the two series from 50 and 49 patients.

Thus, estimated variance due to surgeon $= 5.78 \times 10^{-4}$

However this estimated variance will be partly due to between patient variance. Variance of proportions due to patient variation is

$$0.488 \times (1- 0.488) \times \left(\frac{1}{50} + \frac{1}{49}\right) = 0.01$$

Since this exceeds the estimated variance due to surgeon, we must assume there is no true surgeon effect at all for Prosthesis A.

Prosthesis D; proportions were 0.54, 0.393 with 169 and 83 patients.

Estimated variance due to surgeon $= 108.044 \times 10^{-4}$

Variance due to patients $= 0.4665 \times (1 - 0.4665) \times \left(\frac{1}{169} + \frac{1}{83}\right)$

$= 44.711 \times 10^{-4}$

So true between surgeon variance $= (108.044 - 44.711) \times 10^{-4}$

$= 63.333 \times 10^{-4}$

Prosthesis E; proportions were 0.574, 0.554, with 199 and 39 patients.

Estimated variance due to surgeon $= 2.0 \times 10^{-4}$

Variance due to patients $= 0.564 \times (1 - 0.564) \times \left(\frac{1}{199} + \frac{1}{39}\right)$

$= 75.409 \times 10^{-4}$

Again we must assume the surgeon effect for prosthesis E is 0.

Taking the average of these three estimated variances gives 21.111×10^{-4} (obviously very influenced by results for prosthesis D)

If m is the number of surgeons participating in a trial, the variance of the differences in proportions of good outcomes will be:

$$\text{Variance due to surgeon} \times \left(\frac{2}{m} + \frac{2}{m}\right) = 21.111 \times 10^{-4} \times \frac{4}{m} \quad (2)$$

To avoid increasing the sample size of 525 patients per treatment group, we must equate expressions (1) and (2). This leads to $m=8.87$.

Thus four or five surgeons per treatment group should be used, and they should carry out knee replacements on 100 to 130 patients each.